Objective Examination Methods in the Social Studies

—

RUCH
and
DE GRAFF · GORDON · McGREGOR
MAUPIN · MURDOCK

UNIVERSITY OF
CHARLESTON LIBRARY
1700036255

MORRIS HARVEY COLLEGE
CHARLESTON, WEST VIRGINIA

ANDREW S. THOMAS
MEMORIAL LIBRARY

CALL NUMBER
375.3
R820

ACCESSION
6502

Date Due			
DEC 16 1930			
5/25/31 Resume			
3/23/32 Pres. Burton			
MAY 17 1933			
Pres. Walton 10/15/33			
AUG 22 1934			
MAY 20 '59			
JUN 22 '61			
LOAN AB '65			
MAY 22 '67			

Do not return to library.

Objective Examination Methods in the Social Studies

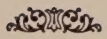
A REPORT OF INVESTIGATIONS CARRIED OUT UNDER A SUBVENTION FROM THE NEW YORK COMMONWEALTH FUND.

by

G. M. RUCH
PROFESSOR OF EDUCATION, UNIVERSITY OF CALIFORNIA
FORMERLY ASSOCIATE PROFESSOR OF EDUCATION AND PSYCHOLOGY
STATE UNIVERSITY OF IOWA

and

MARK H. DE GRAFF
W. E. GORDON
J. B. McGREGOR
NELL MAUPIN
JOHN R. MURDOCK



SCOTT, FORESMAN AND COMPANY
CHICAGO ATLANTA NEW YORK

Copyright, 1926,
by
SCOTT, FORESMAN AND COMPANY

2612.1

CONTENTS

		PAGE
CHAPTER I.	The Purpose and Scope of the Investigations	1
CHAPTER II.	Studies on the Reliability of State Eighth-Grade Examinations	6
CHAPTER III.	Studies on the Reliability of the New York Regents' Examinations	23
CHAPTER IV.	Studies on the Relative Merits of Recall, Multiple-Response, and True-False Techniques with Special Reference to Corrections for Chance	54
CHAPTER V.	A Study of the Technique of Matching Tests	89
CHAPTER VI.	Critical Studies of the Standardized Tests in the Social Studies for High Schools	105
APPENDIX		113

6502

LIST OF TABLES

NUMBER		PAGE
I.	Reliability Coefficients of 16 State Diploma Examinations	9
II.	Average Scores Assigned by Two Different Scorers	10
III.	Differences of Average Scores (of Table II)	11
IV.	Standard Deviations of Marks Assigned by Two Different Scorers	12
V.	Reliability Coefficients of Eight New York Regents' Examinations	35
VI.	Reliability Coefficients of Eight New York Regents' Examinations after Conversion into Objective Form	36
VII.	Differences in Mean Scores and Standard Deviations of Regents' Examinations	37
VIII.	Amounts of Time Used by Pupils in Writing Examinations, both Subjective and Objective Forms (Regular Regents' Examinations)	38
IX.	Amounts of Time Needed for Correction of Examinations, Subjective and Objective	39
X.	Coefficients of Alienation for Mean Values of r in Table V	41
XI.	Numbers of Pupils Used ("Guess" and "Do not Guess" Experiment)	67
XII.	Comparison of Correlations: "Guess" Instructions, Recall A vs. Recognition A	69
XIII.	Comparison of Correlations: "Do not Guess" Instructions, Recall A vs. Recognition A	69
XIV.	Comparison of Correlations: "Guess" Instructions, Recall B vs. Recognition B	70
XV.	Comparison of Correlations: "Do not Guess" Instructions, Recall B vs. Recognition B	70

NUMBER PAGE

XVI. Comparison of Correlations: "Guess" Instructions, Recognition A vs. Recognition B 71

XVII. Comparison of Correlations: "Do not Guess" Instructions, Recognition A vs. Recognition B 71

XVIII. Comparison of Correlations: Uncorrected for Chance. Recall A vs. Recognition B 72

XIX. Comparison of Correlations: Corrected for Chance. Recall A vs. Recognition A 72

XX. Comparison of Correlations: Uncorrected for Chance. Recall B vs. Recognition B 72

XXI. Comparison of Correlations: Corrected for Chance. Recall B vs. Recognition B 73

XXII. Comparison of Correlations: Uncorrected for Chance. Recognition A vs. Recognition B 73

XXIII. Comparison of Correlations: Corrected for Chance. Recognition A vs. Recognition B 73

XXIV. Intercorrelations, Corrected and Uncorrected for Chance, for all Ten Tests 75

XXV. Correlation of Recall and Recognition when Corrected for Attenuation 77

XXVI. Correlations of Recognition A vs. Recognition B, Corrected by Spearman-Brown Formula 78

XXVII. Mean Scores 79

XXVIII. Sigmas 80

XXIX. Percentiles of Time in Minutes for Recall Tests: 100 Items 81

XXX. Percentiles of Time in Minutes for Recognition Types: 200 Items 81

XXXI. Mean Times in Minutes 82

XXXII. Comparisons with Studies of Toops and Ruch-Stoddard 83

XXXIII. Number of Items per Unit of 100 Recall Items 83

NUMBER PAGE

XXXIV. Number of Items of Each Type per Unit of Time for 100 Recall Items 84

XXXV. Analysis of Validity of Corrections for Chance 87

XXXVI. Results for Eighth-Grade Pupils on Series I, Tests 1 to 6 97

XXXVII. Results for High-School Seniors on Series I, Tests 1 to 6 98

XXXVIII. Results for Series I, Tests 1 to 6, for All Pupils 98

XXXIX. Results for Eighth-Grade Pupils on Series II, Tests 7 to 12 99

XL. Results for High-School Seniors on Series II, Tests 7 to 12 99

XLI. Results for Series II, Tests 7 to 12, for All Pupils 100

XLII. Reliability Coefficients for Ten Test Forms 107

XLIII. Intercorrelations of the Ten Test Forms 109

XLIV. Intercorrelations Corrected for Attenuation 110

A. Coefficients of Alienation and the Reduction in the Standard Errors of Estimates for Various Values of r .. 114

OBJECTIVE EXAMINATION METHODS IN THE SOCIAL STUDIES

CHAPTER I

THE PURPOSE AND SCOPE OF THE INVESTIGATIONS

Introduction. The investigations described in this volume were undertaken in order to gather additional information concerning certain of the alleged defects of the traditional examination practices in the social studies and to study critically the claims to superiority of various newly proposed *objective* examination techniques.

Largely through the efforts of Professor Ernest Horn of the State University of Iowa, the New York Commonwealth Fund granted the director of this study the sum of $2500 to cover the expenses of the investigation. The actual experimentation was begun in October, 1924, and was completed in June, 1925.

Some idea of the scope of the experimentation will be had by the following brief statements. A total of 40 different booklets of test materials was prepared and used in the investigations, exclusive of about 100 pages of mimeographed examination questions used in the study described in Chapter II. The total number of pages of printed test materials making up the 40 booklets was 332 pages. All printed materials were set up in 10-point type on 8½ x 11-inch pages. High-grade bond printing paper was used throughout. Every attempt was made to make the printed materials attractive in appearance so as to appeal to the interest of the pupils. In all, 8946 pupils actually took part in one or more experiments, the total pupil-working-time aggregating about 600,000 minutes, or 10,000 hours.

The scoring of the test materials was all done at Iowa City. In the case of the objective tests the scoring was done by clerks using prepared stencils. The readers of the papers requiring trained scorers were invariably experienced teachers who were carefully selected upon the basis of their academic preparation in the social studies and their teaching experience. Moreover, every teacher acting as a reader in any experiment had first studied for from three to six weeks the existing literature on the sources of error in marking school examinations. This was done as a further safeguard to the competency of the scorers of the examinations. It is probably safe to assume that, because of the precautions which have been enumerated, the readers of all experimental examinations constituted a highly selected group. Where the traditional examination practices have been shown to be subject to large errors, the defects can safely be assumed to be inherent in the examinations themselves rather than eliminable effects which were functions of the particular group of scorers actually employed.

The investigational staff included the following: Mark H. DeGraff, Ph. D., Walter E. Gordon, Ph. D., Jay B. MacGregor, M. A., Nell Maupin, M. A., John R. Murdock, M. A., G. M. Ruch, Ph. D. (Director).

Acknowledgments are also due to Dr. Ernest Horn, Dr. Bessie L. Pierce, Dr. H. G. Plum, Miss Clara M. Daley, and Mr. Ralph E. Turner for valuable criticisms of the test materials and suggestions for the conduct of the experiments. Mr. Murdock and Miss Maupin shared in the shaping of the general plans for the investigations. To these should be credited the preparation of most of the experimental materials not specifically acknowledged as to authorship in the chapters that follow.

It was the original intention to make specific mention of the aid rendered by each teacher, principal, or superintendent coöperating in this study. However, it was found that such a list would run to a number of printed pages.

In view of the space limitations the director of the study was forced to omit individual acknowledgments. The willingness of the public schools of the country to place their facilities at the disposal of the investigation was a noteworthy example of the progressive spirit of American education. It is a matter of real concern to the writer that more specific acknowledgment cannot be made to the hundreds of persons who made the present studies possible.

Objectives of the study. In determining the exact direction which these investigations should take, several possibilities presented themselves. One of these which was considered and finally rejected was that of preparing "model" tests and examinations which teachers of the social studies might adopt bodily, adapt to local needs, or imitate in the preparation of examination materials. After considerable thought concerning and study of the present situation with respect to the social-science curriculum, it was decided that such a procedure would be unwise in view of the fact that the courses of study in history and the other social studies are today undergoing very rapid metamorphoses. The preparation of a series of social-science tests and examinations might, as a matter of sheer availability of such materials, tend to perpetuate the traditional curriculum in such subjects in the face of a number of very important curricular studies in progress at the present time, notably the investigations of Dr. H. O. Rugg and his associates.

For these and other reasons it was finally decided to limit the present investigations to two general types, viz., (1) studies of the merits of present examination practices in the social studies, including standardized tests now available, and (2) studies of the relative merits of such objective-examination techniques as matching tests, completion tests, multiple-response tests, and true-false tests, etc. In this connection it is essential that the reader of these pages should bear constantly in mind that *no claims are put forth relative to the validity of the actual test materials except*

to the extent that all materials actually employed may safely be assumed to be equal in quality to those likely to be prepared and used by progressive teachers of more than usual experience with examination practices. The position was taken that, on the whole, the fairest comparisons of the older and newer examination methods would be obtained by exercising no appreciably greater care in the preparation of the experimental tests than would probably be the case with any of the well known series of official examinations, such as the examinations of the College Entrance Board, the Examinations of the University of the State of New York (The Regents' Examinations), or the examinations set by the various states under authority of the state superintendents of public instruction.

Consequently, the experimental tests of this investigation were subjected to no rigorous validation other than the use of the pooled judgments of from two to five "competent" persons, invariably teachers of the social studies or members of the investigational staff. Rigid validation of materials would have been objectionable by virtue of the fact that the resulting tests and examinations would have been artificial in the sense that average teachers could not hope to duplicate the results of the present studies by the use of tests prepared under actual teaching conditions.

All of the experimental tests employed are therefore to be looked upon as *means of evaluating the techniques of testing* rather than as model materials for the guidance of teachers in constructing examinations. Similarly, the aim of the studies was that of *testing the tests themselves;* never that of testing the pupils except as the latter was necessarily involved in the former. Moreover, in the case of certain examinations like those of the several state departments of public instruction or the New York Regents, it should be borne in mind that the conditions under which such materials were used in the present investigations were often widely different from the conditions for which these official

examinations were planned for use. The justification for the departures rests upon the fact that the present investigations were not concerned with the merits of the particular examinations employed, but rather with the *type* of examinations usually adopted by such examining bodies. In fact, in the opinion of the writer many of the official examinations represent the practical limits of perfection along the line of their historical development. Their faults are mainly those inherent in the fallibilities of human judgment, i. e., the impossibility of objective evaluation of pupils' written answers to the questions set. It is hoped that readers will view the results set forth in these pages with the before-mentioned qualifications in mind.

CHAPTER II

STUDIES ON THE RELIABILITY OF STATE EIGHTH-GRADE EXAMINATIONS[1]

Introduction and statement of the problem. Requests were sent to all state superintendents of public instruction for copies of all official state eighth-grade examinations for as many years past as possible. Eleven states responded with questions which were actually used in this investigation. A few other states delayed their returns until too late for inclusion.

The examination questions from the eleven states were classified in three groups: United States history, geography, and civics (citizenship). Key numbers were assigned to the examinations in order that the source of the questions would not be revealed to the schools coöperating in the experiment. These key numbers are used in the tabulations to be presented in this chapter. Every attempt was made to avoid any publicity about the particular states furnishing the questions, since it was the intention of the investigators to study the eighth-grade examination system as a whole rather than to direct attention to the examination practices of particular states. Parenthetically, it may be said that no evidence was found which suggested that any of the individual states were measurably superior or inferior to the others in the character of the examinations set for their eighth-grade pupils.

Occasional questions were omitted when such questions were based upon local history, geography, or government, for two reasons: (1) in order not to reveal the source of the examination, and (2) because the inclusion of such

[1]An abstract of an M. A. thesis presented to the Graduate College of the State University of Iowa by Jay B. McGregor under the title, *A Statistical Study of Eighth-Grade Diploma Examinations*, 1925.

questions would not be a valid procedure in view of the fact that the questions would be used in states other than the one for which the examination was devised. Since the examinations usually offered some degree of choice in the questions to be answered, it was possible to omit an occasional question without much violence to the examination.

The sets of questions were then mimeographed so that each pupil might have his individual copy. The following directions were given to the pupils:

> You are to be given an examination in . . . . . (the teacher supplied the subject), which pupils in another state had to take in order to get their eighth-grade diplomas. We want to see if you can do as well as pupils in other states. Work as fast as you can without making mistakes. When you have finished, record your time in a square which you should make on the last page near the bottom."

The teacher timed the examinations to the nearest one-half minute by means of the plan of writing the elapsed time at half-minute intervals on the blackboard.

All of the pupils used in the experiment wrote on two examinations for the same subject, viz., the set of questions for the year 1923 and the set for 1924.

The examination papers were returned to the University of Iowa for scoring. Thirty-two experienced teachers of the social studies did the scoring, *every paper being marked independently by two teachers.*

That the investigation included a wide sampling of state examinations, pupils, and scorers is shown by the following facts:

(1) The eighth-grade examinations were drawn from eleven different states.

(2) Thirty-two different sets of questions were used, i. e., both the 1923 and 1924 questions for sixteen school subjects.

(3) Thirty-two different teachers read the papers, each teacher reading the 1923 and 1924 examinations for one class of pupils.

(4) The papers include two examinations each from 952 pupils representing 15 schools and 11 states.

All papers were graded upon a basis of 100 per cent. If the examination included 10 questions, each question was allowed a maximum of 10 points. Where 5, 8, etc., questions were employed, the 100 points were divided evenly among the questions.

Treatment of the results. The sixteen examinations permitted the calculation of a total number of 96 correlations, each correlation being a *reliability coefficient* from some point of view. These reliability coefficients fall into three more or less distinct classes, as will be brought out later.

The six correlations possible for each set of examinations may be shown by the following outline:

(1) 1923 examination: scorer No. 1 vs. scorer No. 2.
(2) 1924 examination: scorer No. 1 vs. scorer No. 2.
(3) Scorer No. 1: 1923 examination vs. 1924 examination.
(4) Scorer No. 2: 1923 examination vs. 1924 examination.
(5) 1923 examination scored by No. 1 vs. 1924 examination scored by No. 2.
(6) 1924 examination scored by No. 1 vs. 1923 examination scored by No. 2.

The six numbered columns of Table I correspond to the above numbering scheme.

Table I presents the 96 reliability coefficients possible for the 16 sets of examinations.

TABLE I*

RELIABILITY COEFFICIENTS OF 16 STATE DIPLOMA EXAMINATIONS, YEAR (1923) AGAINST YEAR (1924) AND SCORER AGAINST SCORER.

No.	KEY	SUBJECT	(1)	(2)	(3)	(4)	(5)	(6)	POP.
1	G-2	Ele. Citizenship..	.45	.21	−.05	.46	.34	−.26	102
2	I-1	U. S. History....	.60	.43	.16	.41	.23	.17	31
3	J-1	U. S. History....	.47	.30	.44	.73	.25	.22	32
4	F-3	Geography.....	.58	.39	.37	.22	.17	.55	36
5	F-1	U. S. History....	.89	.99	.67	.64	.67	.69	94
6	D-2	Civics.....	.82	.88	.22	.25	.33	.23	36
7	I-3	Geography.....	.40	.88	.32	.48	.29	.41	32
8	M-2	Civics.....	.80	.82	.47	.55	.46	.52	61
9	D-1	U. S. History....	.81	.22	.54	.65	.49	.35	34
10	L-1	U. S. History....	.79	.57	.73	.48	.45	.66	107
11	A-1	U. S. History....	.81	.85	.66	.71	.56	.65	42
12	B-1	U. S. History....	.53	.58	.36	.34	.27	.41	97
13	B-2	Civics.....	.63	.20	.36	−.18	−.06	.25	82
14	K-1	U. S. History....	.93	.91	.56	.67	.68	.51	99
15	I-2	Civics.....	.81	.53	.37	.27	.52	.46	35
16	E-1	U. S. History....	.75	.12	.26	.59	.60	.19	32
Averages.....			.69	.56	.40	.45	.39	.38	(952)
Averages by pairs of columns.....			.62		.43		.38		

*COLUMN (1) 1923 examination, scorer No. 1 vs. scorer No. 2.
COLUMN (2) 1924 examination, scorer No. 1 vs. scorer No. 2.
COLUMN (3) Scorer No. 1, 1923 examination vs. 1924 examination.
COLUMN (4) Scorer No. 2, 1923 examination vs. 1924 examination.
COLUMN (5) 1923 examination scored by No. 1 vs. the 1924 examination scored by No. 2.
COLUMN (6) 1924 examination scored by No. 1 vs. the 1923 examination scored by No. 2.

Table II shows the average scores or marks assigned to both 1923 and 1924 examinations of the 16 state diploma examinations.

TABLE II

THE AVERAGE SCORES (MARKS) ASSIGNED BY TWO DIFFERENT SCORERS FOR BOTH THE 1923 AND 1924 EXAMINATIONS (THE 16 STATE DIPLOMA EXAMINATIONS).

No.	KEY NO.	(1) 1923 EXAM. SCORER 1	(2) 1923 EXAM. SCORER 2	(3) 1924 EXAM. SCORER 1	(4) 1924 EXAM. SCORER 2
1	G-2	67.5	82.0	73.0	70.4
2	I-1	71.7	67.1	57.0	71.0
3	J-1	68.1	43.6	64.9	45.6
4	F-3	70.0	54.8	70.3	69.9
5	F-1	47.7	45.3	51.4	41.9
6	D-2	55.7	47.5	65.6	59.1
7	I-3	51.0	62.3	50.4	68.9
8	M-2	51.0	48.6	48.5	42.9
9	D-1	38.3	34.4	48.5	30.7
10	L-1	49.3	56.5	38.3	65.3
11	A-1	42.5	28.1	25.3	18.6
12	B-1	14.4	7.7	24.4	25.3
13	B-2	29.3	26.0	24.8	11.5
14	K-1	48.1	59.0	61.3	64.7
15	I-2	38.3	41.4	68.1	58.0
16	E-1	21.0	26.9	8.6	12.4

SUMMARY OF DIFFERENCES[1]

	(1 −2)	(3 −4)	(1 −3)	(2 −4)	(1 −4)	(2 −3)
Average Difference...	8.6	9.4	9.2	9.9	11.4	12.0
Largest Difference....	24.5	17.6	29.8	27.0	23.9	26.8
Smallest Difference...	2.4	2.0	0.3	0.4	0.1	0.2

Table III shows the differences in the average scores (of Table II) assigned by two different scorers for both forms of the 16 state diploma examinations. Algebraic signs are ignored.

The outline at the top of the next page is necessary in

[1](1 −2), (3 −4), etc., refer to the differences in the columns numbered 1, 2, 3, and 4.

interpreting the meanings of the columns lettered (*a*), (*b*), (*c*), etc., in Table III.

(*a*) Differences in the average scores assigned to the 1923 and 1924 examinations by scorer No. 1.

(*b*) Differences in the average scores assigned to the 1923 and 1924 examinations by scorer No. 2.

(*c*) Differences in the average scores assigned to the 1923 examinations by scorers Nos. 1 and 2.

(*d*) Differences in the average scores assigned to the 1924 examination by scorers Nos. 1 and 2.

(*e*) Differences in the average scores assigned when scorer No. 1 read the 1923 examination and scorer No. 2 read the 1924 examination.

(*f*) Differences in the average scores assigned when scorer No. 1 read the 1924 examination and scorer No. 2 read the 1923 examination.

TABLE III

DIFFERENCES IN THE AVERAGE SCORES (OF TABLE II) ASSIGNED BY TWO DIFFERENT SCORERS FOR BOTH FORMS OF THE 16 STATE DIPLOMA EXAMINATIONS. ALGEBRAIC SIGNS ARE IGNORED.[1]

No.	KEY NO.	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)
1	G-2	5.5	11.6	14.6	2.5	2.9	9.1
2	I-1	14.7	3.9	4.6	14.0	0.7	10.1
3	J-1	3.2	2.0	24.5	19.3	22.5	21.4
4	F-3	0.3	15.1	15.2	0.4	0.1	15.5
5	F-1	3.6	3.3	2.5	9.4	5.8	6.1
6	D-2	9.9	11.6	8.2	6.5	3.4	18.1
7	I-3	0.6	6.6	11.3	18.5	17.9	11.9
8	M-2	2.5	5.7	2.4	5.6	8.1	0.2
9	D-1	10.1	3.7	4.0	17.8	7.6	14.1
10	L-1	11.0	8.8	7.2	27.0	16.0	18.1
11	A-1	16.7	9.4	14.4	7.2	23.9	2.3
12	B-1	10.1	17.7	6.7	0.9	10.9	16.8
13	B-2	4.5	14.5	3.3	13.3	17.8	1.2
14	K-1	13.2	5.8	10.8	3.4	16.6	2.4
15	I-2	29.8	16.6	3.0	10.1	19.7	26.8
16	E-1	12.3	14.5	5.9	3.7	8.6	18.2
Averages		9.2	9.4	8.6	9.9	11.4	12.0
Averages by pairs of columns...		9.3		9.2		11.7	

[1]See Table I for subjects involved and numbers of cases.

Table IV shows the standard deviations (variability) of the scores (marks) assigned by two different scorers for both forms (1923 and 1924) of the 16 state diploma examinations.

TABLE IV

STANDARD DEVIATIONS OF THE MARKS ASSIGNED BY TWO DIFFERENT SCORERS FOR BOTH FORMS (1923 AND 1924) OF THE 16 STATE DIPLOMA EXAMINATIONS.

No.	KEY No.	SCORER 1 1923 EXAM.	SCORER 2 1923 EXAM.	SCORER 1 1924 EXAM.	SCORER 2 1924 EXAM.
1	G-2	14.2	10.8	15.5	15.8
2	I-1	13.7	16.9	14.8	11.8
3	J-1	13.7	14.7	11.4	13.6
4	F-3	10.9	13.8	11.4	13.8
5	F-1	17.9	17.4	19.9	17.6
6	D-2	14.7	13.4	11.2	12.3
7	I-3	14.4	10.3	11.0	14.2
8	M-2	13.6	11.6	15.1	14.7
9	D-1	18.0	15.6	17.7	15.2
10	L-1	16.9	19.0	17.2	13.4
11	A-1	16.6	17.0	17.0	11.2
12	B-1	13.7	9.4	12.2	13.1
13	B-2	14.6	14.9	13.3	11.2
14	K-1	20.2	17.2	18.4	17.7
15	I-2	13.3	13.4	12.6	9.7
16	E-1	11.6	13.6	4.4	5.9

Discussion of the results. The significance of Tables I to IV can best be introduced by a brief discussion of the sources of unreliability in written examinations. Unreliability arises from two principal causes: (1) Errors due to limited sampling. (2) Errors due to subjectivity of scoring (the so-called personal equation of the scorer or marker of the papers).

Examination practices at the present time make use of two more or less distinct theories of sampling. The first of these may be called the *intensive* sampling and is represented by the traditional "essay" type of examination. The second, or *extensive* sampling, is characteristic of the more recent "new-type" or "objective" examinations. The

former examination usually consists of five or ten questions which are to be answered exhaustively. The latter is more likely to comprise from 50 to 250 or more narrow questions which sample widely but not intensively.

It is not always borne in mind that any examination is at best a limited sampling of the total field which might be covered by the examination. Testing is therefore invariably partial, never complete. Unreliability due to sampling can be reduced by increasing the number of questions asked. Theoretically, an examination is perfectly reliable only when the sampling is infinitely long.

The two theories of sampling as applied to examinations may be illustrated by the following scheme:

Let *A*, *B*, *C*, *D*, . . . *N* below represent the topics in a particular school subject. Each should be suitable for one broad question of the usual type. Let 1, 2, 3, . . . *n*, represent single facts or items of information, etc. falling under each of the topics denoted by capital letters. Thus:

A	*B*	*C*	*D*	.	.	.	*N*
1	1	1	1	.	.	.	1
2	2	2	2	.	.	.	2
3	3	3	3	.	.	.	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
n	*n*	*n*	*n*	*n*	*n*	*n*	*n*

It is logical to suppose (and it can be shown experimentally[1]) that knowledge of item or question *A*1 is much more likely to guarantee knowledge of items *A*2 and *A*3 than it is to guarantee knowledge of items *B*1, *C*6, and *Xn*. This is merely equivalent to saying that the intercorrelations are higher among items of the *same* column than between items drawn from *different* columns.

[1]For example, in standard tests, like the *Thorndike-McCall Reading Test*, it can be shown that the items based upon the same reading paragraph are more highly interrelated than are items of two different reading paragraphs. Here the paragraphs are analogous to the capital letters in the scheme, and the questions or items based on the paragraphs are analogous to the 1, 2, 3, etc., falling under each capital letter.

The traditional or intensive type of examination tends to the position of including a few (5-10) columns or topics, these being answered in great detail. The newer objective examination tends to the extensive sampling, i. e., a few narrow items drawn from many columns.

A priori, the advantage lies with the extensive sampling as far as reliability of sampling is concerned, since such samples are not as greatly affected by occasional faulty questions, the missing of work due to absence from school, and other obvious factors. It might also be pointed out that the situation with respect to subjectivity of scoring is similar, since the narrow question is less subject to personal opinion than the broader type of question.

The foregoing discussion of the theory of sampling as applied to examinations may seem unduly theoretical to the reader. The application to the investigations summarized in Tables I to IV is, however, quite direct. It may be laid down as a rule that if examinations given year after year by an educational body are to be held valid, *the examination marks must be stable and reliable quantitative measures from year to year.* If a mark of 85 or 92 is to have any real meaning, the examination marks from year to year must meet approximately the following conditions:

(1) The examinations from year to year must yield approximately the same average scores if given to the *same* group of pupils, or to pupils of the same average ability.

(2) The examinations from year to year must yield the same variability of marks (e. g., standard deviations) if given to the *same* pupils, or to groups of pupils of equal abilities.

(3) The marks or scores of individual pupils must remain the same or approximately the same regardless of which year's questions were used.

Although any examination practice must be judged in the light of the foregoing criteria, it should be pointed out

that the use of the traditional 100 per cent grading scale increases the difficulty of meeting these three conditions. Grading methods which involve the assignment of marks upon the basis of rank-orders of the pupils rather than upon per cents have somewhat greater latitude, since it is considerably easier to arrange examinations from year to year which would yield approximately the same rank positions to the same individual pupils than it is to build examinations which will yield the same per cent scores upon a scale of 100 per cent.

The experiments described in this chapter were arranged to test the approximation of state eighth-grade examinations to the ideal conditions we have described. For this reason the examinations set for both the years 1923 and 1924 in a given state and for a given school subject were administered to the same pupils. Table I presents the agreement of the marks earned by the same pupils on the examinations for two successive years in terms of coefficients of correlation. These serve as general measures of the degree to which the third mentioned criterion or condition is being met.

The six columns in Table I may be grouped as three general situations with respect to examination practices:

Situation I. The examination is the constant and the scorer is the variable. This situation allows the factor of subjectivity of scoring to be the main variable, since two independent readers mark the *same* papers. Theoretically, aside from the subjective element in the marks, all correlations of this type would be unity (1.00). Columns (1) and (2) of Table I present the actual coefficients.

Situation II. The scorer is the constant and the examination is the variable. In this situation the factor of subjectivity still enters, but in its minimum effect, since the *same* scorer reads both sets of examinations. Such correlations, theoretically, must always be less than unity, since unreliability due to small samplings (5 to 10 questions) still enters in addition to unreliability due to whatever degree of subjectivity is non-eliminable even when the scorer is constant. We will expect correlations in Situation II to tend to be smaller than for Situation I. Columns (3) and (4) present the actual coefficients.

Situation III. Both the scorer and the examination are variables, the influence of both being at a maximum (within the limits of the present investigation). Here the 1923 examination was read by scorer No. 1 and the 1924 examination by scorer No. 2. The coefficients yielded by this situation will, theoretically, be smaller than in either of the two preceding situations, since it combines the sources of error present in both. Columns (5) and (6) present the actual coefficients.

In passing, it should be pointed out that the actual coefficients obtained are in harmony with the foregoing predictions as to relative magnitudes, the mean values being .62, .43, and .38 for Situations I, II, and III, respectively.

Situation I. That this situation *might* arise in actual examination practices can easily be shown. Suppose that State Y appointed Miss A to read the papers for the year 1923. Assume further that Miss A was taken ill and Miss B was appointed as her substitute. The amount of influence which this change of readers might have had on the marks of the pupils is measured in a general way by the departure of a correlation of .62 from perfect reliability. (See columns (1) and (2) of Table I.) Table *A* of the Appendix shows that the degree of resemblance represented by a correlation of .62 is *roughly 20 per cent better than that arising from the assignment of both sets of marks by pure chance.*

As has been pointed out, aside from subjective disagreement in the scoring, all of these correlations would, theoretically, be 1.00. They might actually be unity if an objective examination were scored by two independent readers.

Situation I probably does not occur often in actual practice, but it surely does occasionally. There can be little doubt that errors of relatively enormous magnitudes are introduced as a consequence of the "personal equation" of the reader or scorer.

Situation II. That this situation is actually possible may also be shown readily. Suppose that State Y appointed the same reader for two successive years (e. g., 1923 and

1924). Different sets of questions were prepared for each year. Assume further that a pupil, Franklin Brown, was taken ill during June, 1923, and was forced to postpone his examination until June, 1924. In such a case we would have a situation roughly parallel to columns (3) and (4) of Table I. The average reliability coefficient under these conditions was found to be .43. *The agreement represented by a correlation of .43 is approximately 10 per cent better than assignment of marks by pure chance.* (See Table *A* of the Appendix.)

It has been suggested that Situation II will, in general, yield smaller reliability coefficients than Situation I, since the former introduces the effects of limited sampling together with those of subjectivity (although the degree of subjectivity is less here).

The data from columns (3) and (4) establish beyond serious question that present examinations suffer greatly through the limited sampling possible in 5- and 10-question examinations of the so-called "intensive" type.

Situation III. This situation is the typical one in actual practice. The several states or other educational bodies which set examinations almost invariably change both the questions and the reader from year to year. If our hypothetical pupil, Franklin Brown, were taken ill in 1923 and forced to wait until the next regular examination period in June, 1924, both the examination and the reader would have been changed.

The general effects upon pupils' marks under Situation III are represented by average coefficients of .38, *a degree of accuracy roughly* $7\frac{1}{2}$ *per cent better than chance assignment of marks.* (See columns (5) and (6) of Table I and Table *A* of the Appendix.)

It is very difficult to accept the truth of such extreme findings as those of Situation III, so closely parallel as it is to the actual working conditions, but thirty-two different

correlations based upon nearly a thousand pupils and representative of eleven different state-examination series cannot be waived aside. It is true that the figures on reliability coefficients presented in this chapter are slightly lower than those reported in other investigations, *but only slightly*, as is shown by the following summary of a study of examinations by Monroe and Souders.[1] They found the median reliability coefficient for sixty-six examinations to be .65, the range being from −20 to .95, a range almost identical with that of Table I. Monroe and Souders used examinations in subjects other than the social studies as well. Ruch[2] found a range of reliability coefficients of the type falling under Situation I of from .44 to .96, and a range of from .39 to .68 for those falling under Situation II. However, but five different examinations were involved in the latter study.

On the whole, the greatest possible source of error in the results presented in this chapter centers about the giving of many of the examinations to pupils in states other than the ones for which the examinations were constructed. Whether important effects arose from this practice is wholly conjectural except for the fact that many examinations did yield reliabilities of from .80 to .95 or higher under these conditions, showing that certain examinations behaved reliably despite their being given under somewhat irregular conditions. Moreover, any errors introduced by giving examinations of one state to pupils in another would have to be relatively enormous to change any of the interpretations or conclusions presented in this chapter.

Variations in the difficulty of examinations. Table II presents evidence that the fate of a pupil taking one of these examinations is not purely a matter of his actual knowledge of the subject, but that his success or failure on the examination is also to a large extent dependent upon the

[1]W. S. Monroe and L. B. Souders, "The Present Status of Written Examinations and Their Improvement," *University of Illinois Bulletin*, Vol. 21, No. 13.

[2]G. M. Ruch, *The Improvement of the Written Examination*. (Scott, Foresman and Co., 1924), p. 53.

particular lot of questions set for any given year. As the table shows clearly, such differences in difficulty from year to year show a central tendency of about nine points where the same person reads both sets of papers. The differences may rise to from eleven to twelve points, on the average, where one reader scores the 1923 papers and a second reader scores the 1924 papers. Worse still, there are extreme cases where the examinations set for two successive years vary twenty-five or more points (per cent) in average difficulty. (See the summary of differences at the bottom of Table II.)

The inadequacy of the 100 per cent grading system in such situations is almost too apparent for comment. If Table II is at all typical of actual conditions, *whole classes will occasionally be rated as "superior" when they are in reality "mediocre" or even "inferior," and the reverse will also be true at times.* How many of these discrepancies are due to subjective factors and how many to real inequalities in the difficulty of the examinations cannot be stated, nor does it matter greatly, since both sources of difference occur inseparably under actual examination conditions.

Table III furnishes the same facts as Table II but in more direct fashion, since it shows the differences between the means of all columns taken by pairs. For example, column (*e*) of Table III is roughly typical of the examination system in actual practice. The differences in this column were those obtained from columns (1) and (4) of Table II, which represented the situation where one person scores the 1923 examinations and a second person scores the 1924 papers. Of the sixteen differences in column (*e*) of Table III, eight, or 50 per cent, are greater than ten points (per cent). The smallest difference found was 0.1 and the largest was 23.9, the average being 11.4.

Worse still, we have been dealing with *averages of whole sets of marks,* not individual marks. The inequalities and errors in the marks assigned *individual pupils* are even

larger at times than these disagreements in averages, since certain sources of error, like unreliability due to limited sampling and (to some extent) subjectivity, are minimized when averages are determined.

Differences in the variability of marks. Table IV shows the standard deviations of the marks assigned to the various sets of papers. Considerable differences are to be noted here as well, the marks on certain sets of papers being closely distributed around the average, while those of other sets of papers are more spread out. Differences are found in the variabilities of marks assigned by one scorer as against another, and in the marks yielded by one examination as against another. A concrete example will show the significance of differences in variability of assigned marks considered along with differences in average marks assigned.

For the first examination, G-2, we have the following data from Tables I and IV:

1923 EXAMINATION

	Scorer 1	*Scorer 2*
Average	67.5	82.0
Standard Deviation	14.2	10.8
Sum	81.7	92.8

In the above tabulation the S. D. has been added to the average in both cases in order to locate analogous positions in the two distributions of marks. A pupil who received a mark of 82 (81.7 to be exact) on the 1923 examinations as read by scorer 1 would have shown a mastery of the subject-matter equal to that of a second pupil earning a mark of 93 (92.8) on the same questions, providing the second scorer read the paper of the second pupil. Thus, in this situation, 82 equals 93. This is but another evidence of the futility of the 100 per cent marking scale now in vogue. In addition to the numerous theoretical objections which can be urged against the 100 per cent marking scale, the data of Tables

I and IV, taken together, show that the units of such a scale are fluctuating and unstable, being almost as much a function of the reader of the paper as they are of the merit of the pupil's answers.

SUMMARY AND CONCLUSIONS

(1) Sixteen state eighth-grade diploma examinations representing eleven different states yielded average reliability coefficients as follows:

(*a*) 0.62 when the examination was constant and the scorer (reader) was the variable.

(*b*) 0.43 when the scorer was constant and the examination was the variable (1923 vs. 1924 examinations).

(*c*) 0.38 when both the scorer and the examinations were variables.

(2) The agreements represented by correlation coefficients of from .38 to .62 are approximately 7½ per cent to 20 per cent better than chance assignment of marks.

(3) The unreliability of examinations is due to two general causes, viz., limited sampling and subjectivity of scoring.

(4) Unreliability due to subjectivity is completely or almost completely eliminable.

(5) Unreliability due to limited sampling may be minimized by the use of examinations composed of 100 to 250 or more short questions rather than five to ten broad questions, i.e., by adopting an *extensive* plan of sampling rather than an *intensive* sample.

(6) Differences in the average (mean) marks assigned to a set of papers by two independent scorers may be as great as 25 per cent, the central tendency of such differences in averages being found to be about eight to twelve points. Such differences were greatest when both the scorer and the examination were variables from year to year.

(7) The differences mentioned argue against the accuracy of the 100 per cent grading scale. Assignment of marks in

terms of ranks would probably be safer, since it is easier to produce examinations from year to year which would yield approximately the same rank-orders for the same pupils than it is to construct examinations which will yield approximately the same mean scores from year to year.

(8) Large differences in the variability (standard deviations) of the distributions of marks assigned by different scorers to the same papers or by the same scorer to the papers of different years were found. Such differences, coupled with the differences in average scores, often result in a mark of 80 per cent on one examination being statistically equivalent to a mark of 65 per cent or of 95 per cent on the examination set for the same subject the next year.

(9) The present results are fairly in harmony with similar studies, notably those of Monroe and Souders and of Ruch.

(10) The results of this investigation show clearly that the present state eighth-grade examination system is largely invalid. Any reasonable allowance for errors of experimental method cannot offset the unsatisfactory conditions which were found. The only serious departure of the method of the investigation from the conditions of actual examinations seems to be found in the plan of giving the examination to pupils in states other than the ones for which the examinations were designed. The effects of this departure are purely conjectural, there being evidence in many cases, at least, that no especial significance attaches to this irregular procedure.

CHAPTER III

STUDIES ON THE RELIABILITY OF THE NEW YORK REGENTS' EXAMINATIONS[1]

I. THE NEED FOR CRITICAL STUDIES

Introduction. The rapid spread of the use of true-false, multiple-response, matching-test, and completion-exercise types of examination in college, high-school, and elementary-school instruction demands that critical studies of the relative merits of these newer methods be carried out at once in order to determine the relative merits of the traditional and so-called "new-type" or objective examinations.

These critical studies should concern themselves principally with such considerations as the following:

(1) Which types of examinations are the most valid?

(2) Which types are most reliable per unit of actual pupil working-time?

(3) Which types are most rapidly scored?

(4) What is the relative amount of time needed to construct each of the different types of examinations (traditional or new-types)?

(5) Which types challenge the highest degree of interest on the part of pupils?

(6) To what types of subject-matter is each kind of examination best suited?

(7) Which types are best adapted to such specific ends as:

(*a*) measurement of factual knowledge?

(*b*) measurement of ability to organize thought?

(*c*) measurement of ability in language expression?

[1]Extracts from a thesis presented to the Graduate College of the State University of Iowa in partial fulfillment of the requirements for the degree of Doctor of Philosophy, by Walter E. Gordon, 1925, under the title, *A Study of the Reliability of Examinations Based on the New York Regents' Examinations in the Social Studies.*

(*d*) measurement of ability to think?
(*e*) measurement of ability to apply general principles?
Etc.

It is the purpose of this study to set forth additional data on certain of the foregoing questions.

The traditional examination is typically represented by the writing of the answers to five to ten questions like the following:[1]

1. Write on architecture and building among the early Egyptians, touching on (*a*) kinds of structures built, (*b*) general plan and appearance, (*c*) materials used.
2. What geographic features of Greece favored (*a*) the growth of small states, (*b*) commerce. Explain fully in each case.

The new type or objective examination might approach the same information as that involved in question 2 above in some such manner as the following:

Directions: Check (X) **two** items which indicate geographic features of Greece favoring the growth of small states.

1. ..... A mountain range divided the country into an eastern and a western Greece.
2. ..... Mountain ranges and water barriers formed numerous plains, thus dividing the people into distinct communities.
3. ..... Natural barriers fostered a love of freedom and of local independence rather than high centralization.

Etc. until 10 or more multiple-choices are presented.

It is obvious that the first form of statement for question 2 permits a wide variety of answers. These will naturally vary in merit from totally unacceptable responses to answers which are worthy of full credit. It follows as an unavoidable consequence that the personal equation of the scorer of the paper will enter into the determination of the exact numerical mark to be assigned a given answer.

Both statements of question 2 have their specific advantages. In order to present the problems under investigation in the present study more clearly, it may help to list the advantages of each.

[1] 229th High-School Examination, The University of the State of New York, History Major Sequence, Course A, Wednesday, June 20, 1923.

Advantages of the Traditional Examination

(1) Provides opportunity for the pupil to organize his information into logical statements.
(2) Provides practice in written expression.
(3) Provides the opportunity for originality of answer.

Advantages of the Objective Examination

(1) Eliminates the subjectivity of the grading of the answers.
(2) Eliminates the writing of hundreds of words, many of which are repetitions.
(3) Allows the actual scoring of the papers to be done by cheap clerical help.
(4) Decreases the time needed for scoring papers.
(5) Releases time ordinarily devoted to actual writing for thinking about the answers.

The disadvantages of each type of examination might also be set forth as follows:

Disadvantages of the Traditional Examination

(1) The scoring is subjective.
(2) Correction of the papers is very time consuming and must be done by experts.

Disadvantages of the Objective Examination

(1) No opportunity for original thought is provided.
(2) The element of chance (guessing) enters into the responses.
(3) There is a tendency for objective examinations to tend toward purely factual questions. (However, there is no experimental evidence to show that the traditional examination, per se, has shown itself to be relatively more free from this criticism.)

Examination of the foregoing summaries of advantages and disadvantages of each type of examination shows clearly that *the choice of the type of examination to be employed depends upon the specific purpose of the examination.* It is very unlikely that experiment will show either type of examination to be superior for *all* purposes. On the other hand, for any particular item, i.e., for any particular question, some choice of examination technique must be made, since but one form of question can be used for that item in the examination.

The lines along which choice of examination techniques must be made may be shown by the following statement of the "Criteria of a Good Examination" listed by Ruch:[1]

1. Validity.
2. Reliability.
3. Objectivity.
4. Ease of administration and scoring.
5. Standards.

These five criteria are by no means mutually exclusive, and they are far from equal in importance. The first mentioned is by far the most important. However, the validity of an examination depends in part upon its reliability, which in turn depends in part upon its objectivity. Reliability (and hence, indirectly, validity) also depends markedly upon the length of the examination, i.e., the adequacy of the sampling.

II. THE METHOD OF THE INVESTIGATION

The problem defined. The present investigation was planned to study the relative degree of validity possessed by a typical series of the traditional or "essay" type of examinations, when given (*a*) in their regular or official form, and (*b*) when converted, item by item, into the objective form.

[1] G. M. Ruch, *The Improvement of the Written Examination* (Scott, Foresman and Co., 1924), p. 11.

The investigation is further limited to that aspect of validity which is commonly known as *reliability*. Due to the lack of an adequate outside criterion, the more general problem of validity cannot be attacked in the present study. Although this limitation must be admitted frankly, it is nevertheless true that the major controversy in the matter of examination techniques does center about the matter of reliability and its contributing factors, e.g., objectivity and adequacy of sampling.

Several possible sets of examinations were available for the purposes of studies similar to this, notably, the examinations of the College Entrance Board and those of the New York Regents (officially known as the Examinations of the University of the State of New York). The latter were finally selected. No particular reasons can be given for this choice other than the fact that they appear to be typical and of more than the usual degree of merit.

The specific questions which this investigation undertook to answer may be listed as follows:

(1) To what extent would two independent scorings of the papers written on the same examination agree? This is equivalent to obtaining an estimate of the amount of subjectivity or personal equation in scoring *identical* papers.

(2) To what extent would the results of the examinations set for any two consecutive years agree if the subjects and the scorer were kept constant? It will be readily recognized that this is, in the main, the problem of sampling.

(3) To what extent would the results of two sets of examinations agree if the questions of two consecutive years were answered by the same pupils but each set were scored by a different teacher? This would introduce into the problem, simultaneously, both the errors of subjectivity and of sampling.

(4) Which would be the more reliable, the New York Regents' examinations in official form or the same questions restated in objective form?

(5) Which sort of examination (traditional or objective) is most rapidly answered, i.e., most economical of pupils' time? This answer is important, since examinations are generally set for one, two, three, or more hours working time, and the extent of the sampling possible in that length of time is determined in part by the form of the actual questions and the method by which the pupil designates his answers.

(6) Which type of examination is more quickly scored?

(7) Which of the various sources of error in examination practices are most easily eliminated and which are quantitatively the most serious?

The test materials. In all, sixteen eight-page test booklets were prepared and used in this investigation. Eight of these represented the New York Regents' examinations for the years 1923 and 1924 for each of the four-year courses in the social studies for New York high schools. These were reprinted without changes other than the following:

(1) Only one question was selected in each group of two or three from which elections could normally be made. In almost every case the first question in each group was taken.

(2) Instead of printing all questions on one page, one or two questions only were placed on a given page of the booklets in order that plenty of room would be left for the writing of the answers directly below the questions.

For each subject of the four-years' course both the 1923 and 1924 examinations were given to each pupil taking part in the experiments. This was planned as a check upon the variation of the difficulty, sampling, etc., of the examinations from year to year.

Each of the eight examinations was turned into objective form, with as little distortion of the original terminology and purpose of the examination as was consistent with the mechanics of an objective examination.

This set of examinations was labelled "Series V, Tests 25 to 40." The following is a complete list by title:

I

(1) *Series V, Test 25.* History, Major Sequence, Course A, 229th High-School Examination, Wednesday, June 20, 1923. (Subjective Form, i.e., the official examination with the changes listed above)

(2) *Series V, Test 26.* History, Major Sequence, Course A, 230th High-School Examination, Wednesday, January 23, 1924. (Subjective Form)

(3) *Series V, Test 27.* History, Major Sequence, Course A, 229th High-School Examination, Wednesday, June 20, 1923. (Test 25 in objective form)

(4) *Series V, Test 28.* History, Major Sequence, Course A, 230th High-School Examination, Wednesday, June 20, 1924. (Test 26 in objective form)

II

(5) *Series V, Test 29.* History, Major Sequence, Course B, 229th High-School Examination, Wednesday, June 20, 1923. (Subjective Form)

(6) *Series V, Test 30.* History, Major Sequence, Course B, 230th High-School Examination, Wednesday, January 23, 1924. (Subjective Form)

(7) *Series V, Test 31.* History, Major Sequence, Course B, 229th High-School Examination, Wednesday, June 20, 1923. (Test 29 in objective form)

(8) *Series V, Test 32.* History, Major Sequence, Course B, 230th High-School Examination, Wednesday, January 23, 1924. (Test 30 in objective form)

III

(9) *Series V, Test 33.* History, Major Sequence, Course C, 229th High-School Examination, Wednesday, June 20, 1923. (Subjective Form)

(10) *Series V, Test 34.* History, Major Sequence, Course C, 230th High-School Examination, Tuesday, January 22, 1924. (Subjective Form)

(11) *Series V, Test 35.* History, Major Sequence, Course C, 229th High-School Examination, Wednesday, June 20, 1923. (Test 33 in objective form)

(12) *Series V, Test 36.* History, Major Sequence, Course C, 230th High-School Examination, Tuesday, January 22, 1924. (Test 34 in objective form)

IV

(13) *Series V, Test 37.* Civics, 229th High-School Examination, Thursday, June 21, 1923. (Subjective Form)

(14) *Series V, Test 38.* Civics, 230th High-School Examination, Thursday, January 24, 1924. (Subjective Form)

(15) *Series V, Test 39.* Civics, 229th High-School Examination, Thursday, June 21, 1923. (Test 37 in objective form)

(16) *Series V, Test 40.* Civics, 230th High-School Examination, Thursday, January 24, 1924. (Test 38 in objective form)

This arrangement of the examinations in groups of four enables the calculation of reliability coefficients as follows:

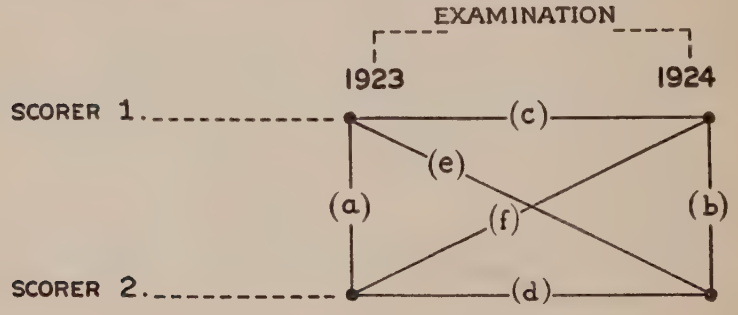
(1) Reliability of two samplings (1923 vs. 1924), scorer being constant. Two such coefficients can be computed.

(2) Reliability of two scorings (two independent scorers), year (either 1923 or 1924 examination) being constant. Two such reliability coefficients can be computed.

(3) The reliability obtained when one sample (e.g., 1923 examination) scored by one scorer is correlated against the second sampling (e.g., the 1924 examination) scored by the second scorer. Two such coefficients can be computed.

(4) The reliability of the two objective examinations when the 1923 sampling is correlated against the 1924 sampling. One such coefficient is possible in each set of four examinations.

Taking any one set of four examinations from the above list of sixteen, it is possible to diagram the six reliability coefficients possible for the subjective examinations.



The correlations of the type (*a*) and (*b*) show mainly the effects of the subjectivity of the scoring, since the same papers are scored in all cases by two scorers working independently.

The correlations of the type (*c*) and (*d*) show mainly the effects of the limited sampling possible from year to year, since the effects of subjectivity are kept at a minimum by having all papers of both years read by the same scorer.

The correlations of the type (*e*) and (*f*) combine both the errors of sampling and the errors of subjectivity, since one year's examination is read by one person and the next year's is read by a second person. Probably this situation is most nearly representative of actual practice.

The above diagram does not show the correlations possible for the two objective examinations (the 1923 vs. the 1924). Only one such correlation is possible for each pair of examinations, and it does not involve errors other than those of limited sampling (the discussion being confined to the types of errors already mentioned).

The method of administration of the examinations. In discussing the administration of these examinations, it should be kept in mind that *each pupil took two examinations.* Moreover, these were invariably the 1923 and 1924 examinations in the same subject, viz., History, Major Sequence, Course A, or Civics, etc.

This is equivalent to saying that the total group of pupils to whom the examinations were given can be considered as breaking themselves into eight sub-groups as follows:

(1) Group I.	Test 35 followed by Test 26.	(Subjective)
(2) Group II.	Test 29 followed by Test 30.	"
(3) Group III.	Test 33 followed by Test 34.	"
(4) Group IV.	Test 37 followed by Test 38.	"
(5) Group V.	Test 27 followed by Test 28.	(Objective)
(6) Group VI.	Test 31 followed by Test 32.	"
(7) Group VII.	Test 35 followed by Test 36.	"
(8) Group VIII.	Test 39 followed by Test 40.	"

The instructions were to give one examination one day and the second examination within a few days of the first. The classes used included classes in ancient history, mediæval and modern history, United States history, and Civics in grades IX to XII.

The total number of test booklets sent out was 7850. These were approximately equally distributed throughout the eight groups which have just been described. A random method of distribution was used to insure that the abilities of the groups taking the subjective and objective forms of the tests could be safely assumed to be approximately equal.

About 5000 blanks were returned to the writer, but because of eliminations through unpaired papers, etc., the final number was reduced to 3766 papers, as follows:

Tests 25 and 26	128
Tests 27 and 28	207
Tests 29 and 30	244
Tests 31 and 32	328
Tests 33 and 34	116
Tests 35 and 36	376
Tests 37 and 38	224
Tests 39 and 40	260
	$1883 \times 2 = 3766$

A wide variety of schools, states, and classes was involved.

The scoring of the examinations. The objective examinations were scored by means of answer keys which gave specific directions for the giving of credit. Certain questions were given rough inspectional weights in order that important items having a small number of possible credits would not be swamped by the larger numbers of credits which might be earned on questions of lesser importance but which, as a matter of convenience or preference, happened to carry a larger number of possible credits. The only criterion of the justice of such weightings used was

reference to the weightings assigned by the original drafters of the Regents' examinations.

It is obvious that more defensible weightings than the ones actually used could have been developed by statistical study of the scores, but this would not have been a fair practice, since the original examinations of the New York Regents were (presumably) weighted by inspection. It follows that no great significance will attach to differences in mean scores or standard deviations between the two paired objective examinations in any of the four subjects. *This, however, does not hold for the subjective examination pairs,* since the weightings actually used by the New York Regents *did determine,* in part, actual success or failure on these examinations. A difference in mean scores or in variability on the original examinations, pair by pair, can genuinely be held to be a disturbing factor in the validity of the examinations.

The scoring of the subjective examinations was carried out in the usual manner, the traditional 100 per cent scale being used. The scorers were in all cases teachers of history with normal, college, or university training in the subject. Moreover, all of these teachers had had three weeks of training in the matter of the "pitfalls" of grading examinations and were as a consequence unusually critical and careful in their grading.

Each scorer graded both the 1923 and the 1924 examinations in the same subject. These two sets of papers were those of the same group, usually 50 pupils. However, the sets of papers for the 1923 examinations were given to the scorers first and were scored and returned by the scorers without knowledge that they were later to grade the 1924 papers. This guaranteed that they had returned to the writer the marks on the 1923 papers before they were given the second assignment. When the 1924 papers were given to the scorers, they were not told that the second set involved the same pupils as the 1923 set previously scored.

It is possible that some of the scorers did discover this fact, but it is doubtful if they could have recalled the marks previously given to the individual pupils. However, it should be noted that if such memory effects were present, they would tend to raise the reliability of the examinations rather than to work unfairly against the reliability of the examinations.

A second scorer read both the 1923 and 1924 examinations in the same subject in the manner just described for the first scorer.

III. STATISTICAL TREATMENT OF THE RESULTS[1]

Reliability coefficients. Table V gives the reliability coefficients of the Regents' examinations in original or official form. Table VI does the same for the objective variates of the same examinations.

It should be noted that instead of having two readers read all of both sets of papers for any one examination, the total lot of papers was broken into a number of sub-lots of 50 (or in some cases fewer) papers. This had two important advantages: (1) A larger sampling of readers was thus secured, and (2) the sub-lots are more typical of the numbers and ranges of ability found in typical school classes.

Table VII presents the differences in mean scores and standard deviations of two independent scorers.

Table VIII summarizes the average time used by pupils in writing both subjective and objective examinations.

Table IX gives the time consumed in the correction of both types of examinations.

---

[1]A section of the original manuscript has been deleted here because a similar, although briefer, account was given in Chapter II of the significance and assumptions of the several types of reliability coefficients computed. The omitted sections also treated of the theory of sampling as applied to examinations. The reader may be referred to Chapter II for the essentials of the omitted sections.

TABLE V

Reliability Coefficents of Eight New York Regents' Examinations in the Social Studies, Year vs. Year, and Reader vs. Reader, Subjective Form

Examination	Sub-lot	(a)*	(b)	(c)	(d)	(e)	(f)	N
Major Sequence, Course A, 1923 (Test 25) vs. 1924 (Test 26)	1	.78 ±.040	.73 ±.044	.39 ±.080	.74 ±.043	.41 ±.079	.38 ±.081	50
	2	.42 ±.078	.85 ±.026	.04 ±.095	.47 ±.074	.02 ±.095	.41 ±.080	50
Major Sequence, Course B, 1923 (Test 29) vs. 1924 (Test 30)	1	.93 ±.013	.74 ±.043	.27 ±.088	.76 ±.040	.61 ±.059	.62 ±.058	50
	2	.70 ±.048	.81 ±.034	.53 ±.060	.73 ±.045	.36 ±.083	.83 ±.031	50
	3	.68 ±.051	.67 ±.054	.48 ±.075	.46 ±.076	.60 ±.061	.45 ±.076	50
	4	.80 ±.034	.84 ±.028	.72 ±.045	.74 ±.043	.61 ±.060	.78 ±.037	50
	5	.88 ±.022	.83 ±.032	.59 ±.066	.78 ±.040	.66 ±.057	.49 ±.076	44
Major Sequence, Course C, 1923 (Test 33) vs. 1924 (Test 34)	1	.82 ±.030	.26 ±.089	.24 ±.089	.54 ±.067	.54 ±.067	.16 ±.092	50
	2	.79 ±.035	.86 ±.024	.47 ±.07	.61 ±.059	.46 ±.075	.53 ±.050	50
Civics, 1923 (Test 37) vs. 1924 (Test 38)	1	.81 ±.033	.70 ±.048	.01 ±.095	.18 ±.092	.04 ±.095	.15 ±.093	50
	2	.68 ±.051	.56 ±.065	.24 ±.090	.28 ±.087	.31 ±.086	.36 ±.094	50
	3	.68 ±.050	.56 ±.065	.27 ±.088	.04 ±.095	.11 ±.094	.10 ±.094	50
	4	.67 ±.052	.58 ±.063	.13 ±.094	.30 ±.087	.28 ±.087	.17 ±.092	50
Means of the Columns		.74	.69	.34	.51	.38	.42	
Means of the Pairs of Columns		.72		.42		.40		

*Column (a). r when 1923 examination was read independently by Scorer No. 1 and No. 2.
Column (b). r when 1924 examination was read independently by Scorer No. 1 and No. 2.
Column (c). r when both 1923 and 1924 examinations were read by Scorer No. 1.
Column (d). r when both 1923 and 1924 examinations were read by Scorer No. 2.
Column (e). r when Scorer No. 1 read the 1923 examination and Scorer No. 2 read the 1924 examination.
Column (f). r when Scorer No. 2 read the 1923 examination and Scorer No. 1 read the 1924 examination.

TABLE VI

RELIABILITY COEFFICIENTS OF EIGHT NEW YORK REGENTS' EXAMINATIONS IN THE SOCIAL STUDIES AFTER CONVERSION INTO OBJECTIVE FORM. ALL CORRELATIONS BASED UPON COMPARISONS OF MARKS EARNED ON THE 1923 AND 1924 EXAMINATIONS.

EXAMINATION	r	P. E.	N
Major Sequence, Course A, 1923 (Test 27) vs. 1924 (Test 28)	.71	.023	200
Major Sequence, Course B, 1923 (Test 31) vs. 1924 (Test 32)	.70	.019	328
Major Sequence, Course C, 1923 (Test 35) vs. 1924 (Test 36)	.60	.027	250
Civics, 1923 (Test 39) vs. 1924 (Test 40)	.59	.027	250
Mean r	.65		

Test No.	27	28	31	32	35	36	39	40
Mean	44.6	44.3	39.8	27.8	48.9	48.2	33.0	35.2
S. D.	12.6	13.9	10.8	8.6	14.4	10.8	10.9	11.9

TABLE VII

DIFFERENCES IN THE MEAN SCORES AND IN THE STANDARD DEVIATIONS OF THE REGENTS' EXAMINATIONS, SAMPLE (YEAR) BY SAMPLE AND SCORER AGAINST SCORER (SUBJECTIVE EXAMINATIONS ONLY).

TEST No.	SUB-LOT	N	MEAN		DIFF. M_1-M_2	SIGMA		DIFF. $S.D_1-S.D_2$
			1st Scorer	2nd Scorer		1st Scorer	2nd Scorer	
25	1	50	41.4	41.8	−.4	18.2	16.9	1.3
25	2	50	41.2	43.3	−2.1	13.5	10.2	3.3
26	1	50	28.0	46.6	−18.6	15.5	18.6	−3.1
26	2	50	34.8	39.4	−4.6	10.7	12.7	−2.0
29	1	50	55.1	47.2	7.9	23.6	23.2	0.4
29	2	50	52.8	55.2	−2.4	15.0	19.2	−4.2
29	3	50	42.1	40.5	1.6	18.8	13.6	5.2
29	4	50	50.4	51.0	−.6	20.6	19.6	1.0
29	5	44	52.3	55.9	−3.6	18.4	20.0	−1.6
30	1	50	30.5	48.3	−17.8	14.1	17.0	−2.9
30	2	50	62.3	52.5	9.8	18.7	15.4	3.2
30	3	50	60.5	41.8	18.7	23.5	14.8	8.7
30	4	50	41.4	43.5	−2.1	17.5	13.2	4.3
30	5	44	47.9	46.0	1.9	25.2	18.5	6.7
33	1	50	63.9	52.9	11.0	21.2	17.0	4.2
33	2	50	39.3	35.3	4.0	19.2	17.1	2.2
34	1	50	32.2	47.2	−15.0	16.8	16.5	0.3
34	2	50	37.1	34.0	3.1	22.7	19.4	3.3
37	1	50	65.0	58.3	6.7	19.6	16.4	3.2
37	2	50	72.3	64.4	7.9	10.0	13.7	−3.7
37	3	50	54.5	53.5	1.0	13.0	16.2	−3.2
37	4	50	66.5	63.7	2.8	15.5	17.8	−2.3
38	1	50	52.2	52.6	−.4	17.9	15.8	2.1
38	2	50	59.8	61.1	−1.3	13.7	15.7	−2.0
38	3	50	51.4	53.2	−1.8	15.6	15.5	0.1
38	4	50	52.1	60.7	−8.6	18.5	18.4	0.1
Mean Difference					6.0			2.9
Range of Differences					0.4–18.7			.0.1–8.7

TABLE VIII

SUMMARY OF THE ACTUAL AMOUNTS OF TIME USED BY PUPILS IN WRITING THE EXAMINATIONS, BOTH SUBJECTIVE (REGULAR REGENTS' EXAMINATION) AND OBJECTIVE FORMS

SUBJECTIVE				OBJECTIVE			
Test No.	No. of Papers	Total Time in Minutes	Average Time in Minutes	Test No.	No. of Papers	Total Time in Minutes	Average Time in Minutes
25	128	3380	26.4	27	207	4347	21.0
26	128	3067	24.0	28	207	4036	19.5
29	244	8553	35.0	31	328	10988	33.5
30	244	7853	32.2	32	328	9840	30.0
33	116	4673	40.3	35	376	13724	36.5
34	116	3808	32.8	36	376	12972	34.5
37	224	11692	52.2	39	260	10270	39.5
38	224	9810	43.8	40	260	7280	28.0
TOTALS	1424	52836	37.0		2342	73457	31.4

$$\frac{\text{Average Subjective Time}}{\text{Average Objective Time}} = \frac{37.0}{31.4} = 1.18$$

TABLE IX

SUMMARY OF AMOUNTS OF TIME NEEDED FOR CORRECTION OF EXAMINATIONS, SUBJECTIVE AND OBJECTIVE.

No. Papers	Subjective		Objective	
	Test	Time in Min.	Test	Time in Min.
128	25	347	27	309
128	26	380	38	567
244	29	934	31	1043
244	30	1021	32	849
116	33	553	35	504
116	34	426	36	547
224	37	1008	39	1134
224	38	1260	40	945
Totals.....		5929	.....	5898
Mean.....		4.2	.....	4.1

IV. DISCUSSION OF THE RESULTS

Discussion of Table V. An important source of error in the traditional or "essay" type of written examination is to be found in the *subjectivity of the scoring.* This statement is founded upon the alienation of the mean r of columns (*a*) and (*b*) of Table V from unity. The mean value, .72, represents an alienation of 69 per cent from perfect reliability. (See Table X and Table A of the Appendix.) Since the examination is the constant in both columns and the scorer is the variable, the alienation is *all* attributable to differences in the subjective factors present in the minds of the readers of the papers. (See Situation I of Chapter II.) As was pointed out in Chapter II, such values of r could be as high as .99+ or even 1.00 under purely objective scoring. (See also Table X or Table A of the Appendix for the detailed calculation and evaluation of the r's in terms of the coefficient of alienation.)

A second important source of error in the written examina-

tions under discussion is the *error due to sampling.* This conclusion rests upon the fact that columns (*c*) and (*d*) of Table V show a mean value of r equal to .43, an alienation of 90 per cent from perfect reliability. As was discussed under Situation II of Chapter II, this alienation is due in part to subjectivity and in part to inadequate sampling. It is not possible to state the relative shares of each of the two causes of unreliability, since, in addition to the differences which are caused by lack of perfect relation of 1923 and 1924 examinations as a pure matter of sampling, there are other sources of error (subjective error) arising from the fact that the keeping of the scorer constant does not eliminate all subjective error. It is true that the constant scorer probably reduces the error over the situation of different scorers, but the present data are not adequate to state the relative or absolute amounts of such reduction.

It may be useful to think of the situations covered by columns (*c*) and (*d*) as characterized by the minimum subjective error (in a subjective examination) since the *same* scorer read both sets of papers. Columns (*e*) and (*f*) would, therefore, be characterized by the maximum subjective error, since the scorer (as well as the examination or sampling) is also a variable.

When both the examination (sampling) and the scorer are variables, the alienation from perfect reliability is greater than in either of the two preceding cases, i.e., the mean r was found to be .40 (an alienation of 92 per cent), but the difference is not very great between the mean values of r in columns (*c*) and (*d*) in comparison with the mean values of r in columns (*e*) and (*f*). (See Situation III of Chapter II.)

It is probably impossible to state whether subjectivity or limited sampling is quantitatively the more important source of error. The only theoretical way (and there is no practical way) of arriving at knowledge of the relative effects of sampling and subjectivity would involve the scoring of the

1923 examination by an *infinite* number of readers, and averaging this infinite series of marks; the 1924 examination would be treated in the same manner. The resulting r would involve no subjective error (as a matter of the foregoing definition) and would represent purely the error of sampling. If this r_∞ were compared with the r's obtained from large numbers of correlations of the types found in columns (*a*) and (*b*) of Table V, the question of the relative influences of sampling and subjectivity could be answered. This, of course, offers no practical help in our evaluation. A practical approach could be had if large numbers, say 100 or more, of readers read both the 1923 and 1924 examinations and the average marks of the 100 teachers were correlated.

TABLE X

COEFFICIENTS OF ALIENATION FOR THE MEAN VALUES OF r IN TABLE V.

r	COLUMNS (MEANS OF)	$\sqrt{1-r^2}$	PER CENT OF REDUCTION IN STANDARD ERROR OF ESTIMATE
1.00		.00	100%
.72	(*a*) and (*b*)	.69	31%
.43	(*c*) and (*d*)	.90	10%
.40	(*e*) and (*f*)	.92	8%

Discussion of Table VI. Table VI shows the reliability coefficients for the objective forms of the same examinations. The r's range from .59 to .71, with a mean value of .65 (alienation 78 per cent). The following conclusions may be drawn from Table VI:

(1) Even when the factor of subjectivity is completely ruled out by means of mechanical scoring rules and objective forms of questioning, the reliability coefficients are not very high. This alienation must be attributed largely to limited sampling. *It must be borne in mind that changing the Regents' examinations to objective form does not materially,*

if at all, increase the range of sampling, since either eight or ten questions were the numbers of questions involved.

(2) The fact that the reliability coefficients of the objective examinations are not high *does not necessarily discredit the claims for objective examinations* which have recently been put forward, since the actual sampling involved in the present examinations does not approximate the theory of extensive sampling which is the heart of the claim of the newer objective examination methods. (See Chapter II.) The present objective examinations are to be thought of as "intensive samplings" over a few units (eight to ten questions) of the subject; the more typical objective examination aims at "extensive sampling" over many units of the subject, the samplings *within* the units being much less detailed.

The present objective examinations are to be thought of as *objective variates* of the Regents' examinations rather than de novo objective examinations over the same subjects. The objective examinations under consideration probably do not materially increase or decrease the range of genuine sampling in comparison with the subjective forms of the same examinations. Certainly, it is true that had the effort been made to produce an objective examination over a year's work in ancient history (rather than "translate" a subjective examination into objective form), a different theory of sampling and unquestionably a wider range of sampling would have been employed. This fact, alone, will explain the lack of agreement of the reliability of the present objective examinations with the objective examinations discussed in Chapter IV.

(3) The mean r (.65) of Table VI is probably most fairly compared with the mean value of r for columns (*e*) and (*f*) of Table V (i.e., .40), since these two sets of reliability coefficients are the most representative of actual examination conditions.

Since both involve limited samplings (1923 vs. 1924 examinations) and the factor of subjectivity is the variable, the difference represents the relative improvement possible through objectifying the pupils' responses. This difference, although not as great as might have been shown had more ideal objective examinations (as defined by the discussion of the theory of sampling in Chapter II) been employed, is nevertheless quite significant. In terms of alienation coefficients, .65 represents an improvement of 14 per cent over .40 in the reduction of the standard error of estimate. The 14 per cent is obtained as follows:

r	$\sqrt{1-r^2}$	Per Cent of Reduction of Standard Error of Estimate
.65	.78	22%
.40	.92	8%
Difference (Improvement)		14%

There can be little doubt that even under the limiting conditions of the objective examinations actually used, the objective examinations were considerably superior in reliability to the subjective forms of the same examinations. Under greater freedom this superiority might be greatly increased.

(4) In comparing the objective and subjective examinations, another factor must be considered, viz., that somewhat unequal amounts of working time were required to answer the two sets of examinations. (See Table VIII.) Since the subjective examinations required, on the average, 1.18 as long a working-time in comparison with the objective examinations, *the only fair comparison would be that of comparing a subjective examination with one 1.18 times as long in actual working time.* This will be done in a later section.

(5) There is a possibility that the comparison of reliabilities of objective and subjective examinations may be dis-

turbed by differences in the range of talent. As is well known, the greater the range of individual differences present in the group upon which the reliability coefficient is figured, the greater the correlation, all other things being equal. There is no method of comparing directly the ranges of talent used in objective and subjective examination groups, since the comparisons ordinarily made by means of the standard deviations assume that the tests were scaled to the same units. This was not true. The subjective examinations were scored upon a 100 per cent scale using the weightings supplied by the original Regents' examinations. The objective examinations were weighted by inspection and often totaled considerably more or less than 100 points.

In terms of score units the standard deviations of the objective examinations were much smaller than those of the subjective. Probably no significance is to be attached to this fact.

There is, however, one item of information about the relative ranges of pupil-talent involved in the two sets of examinations, viz., that the original group was divided into two sub-groups by chance, one group taking the objective and the other the subjective form of a given examination, e.g., Civics. By and large, with total populations as great as 644 and 950, respectively, for subjective and objective examination, differences in range of talent must be practicably negligible, simply as a matter of probabilities.

(6) Attention must be directed to the method of weighting the separate tests of the objective examinations. As has been stated, this was done by inspection; occasionally two or three persons arrived at a particular weighting in conference.

Had careful statistical analysis of such factors as the relative variabilities and reliabilities of the separate parts of the tests been resorted to, there can be little doubt that weightings leading to much higher reliability coefficients could have been produced.

Here, again, the conditions were set so as to limit the objective examinations in a manner not consistent with the theory upon which the claims of the objective examination are based; i.e., typically, the objective examination is not weighted at all, since the sampling is wider and the items less detailed. If but five or ten questions, each fairly broad in scope, are asked, weighting might be a serious problem. If 100 to 500 narrow questions (objective) are asked, the need for weighting disappears as a practical examination problem.

No defense is put forward for the actual weightings used in the objective examinations. *The tests were weighted by inspection, i.e., personal judgment, because it was assumed that such was the method by which the original Regents' examinations were weighted.* On the whole, the weighting of the objective examinations probably has made for fairer comparisons.

Discussion of Table VII. Table VII presents evidence on several detailed factors involved in the evaluation of the merits of examinations, viz., variations in the means and standard deviations from one examination (sampling) to the next. As has been pointed out, such differences arise from (*a*) real differences in the difficulties of the examinations from year to year, and (*b*) differences which are a function of the personal judgments of the readers of the papers.

The following statements summarize the facts of Table VII:

(1) When two different scorers marked the same papers, the mean difference in the marks assigned was found to be 6.0 points on a scale of 100 points.

(2) The largest difference noted in twenty-six sets of papers was 18.7 points, and the smallest difference was 0.4 point.

(3) Six times in the twenty-six sets of papers, the differences were as great as 10 (approximately) points.

(4) In several cases the differences in the mean marks assigned by two independent readers were as large as or larger than the standard deviation of either's distribution of marks. Examples of this situation are: Test 26, sub-lot 1; Test 30, sub-lot 1; Test 30, sub-lot 3.

Using $M_1 - M_2$ as the difference and the S. D.'s as a basis of comparison, we find the following facts for the three cases mentioned immediately above:

Test	Sub-lot	N	M_1	M_2	$M_1 - M_2$	S.D.$_1$	S.D.$_2$	$\frac{M_1 - M_2}{\text{S.D.}_1}$	$\frac{M_1 - M_2}{\text{S.D.}_2}$
26	1	50	28.0	46.0	18.6	15.5	18.6	1.20	1.00
30	1	50	30.5	48.3	17.8	14.1	17.0	1.26	1.05
30	3	50	60.5	41.8	18.7	23.5	14.8	0.80	1.26

If the displacement of the mean marks assigned by two different teachers to the same set of 50 papers is equal to 1.00 or more S. D. in 10 per cent or more of the cases, the injustices arising from the use of such examinations require no lengthy comment. Such inequalities might easily cause 5 per cent of pupils to fail one year's examination and 20 per cent to 25 per cent to fail the next year's examination *without the existence of any differences at all* in the abilities of the two groups of pupils.

As has been mentioned before, these differences are due to two major causes: (*a*) subjectivity of marking, and (*b*) genuine differences in the difficulties of the two examinations due to inadequate sampling.

(5) Similar differences in the standard deviations of the marks of the two independent scorers were found, the range of differences being from 0.1 to 8.7 points with a mean difference of 2.9 points.

(6) In accepting these figures at face value, it is necessary to weigh the possibility of the differences which we have just been discussing because of the fact that the teachers

who scored the twenty-six sets of examinations were not competent readers. It might be objected that the readers of the official Regents' examinations under regular conditions are highly selected and unusually well trained persons, superior as a group to the present group of teachers. Whether this is true or not, really does not matter greatly, since the present investigation is not aimed primarily at the evaluation of the New York Regents' examinations as actually administered. The Regents' examinations were merely used as a convenient and presumably typical set of examinations for revealing sources of error in examinations in general. The conditions under which the Regents' examinations are normally administered are sufficiently different from those of the present study to make evaluations have indirect rather than direct bearing. In this connection it should be pointed out further that the pupils used in this investigation had, in no cases, studied the New York State course of study upon which the regular Regents' examinations are based.

This fact will tend to explain the relatively poor scores earned on the examinations. (Table VII shows that the mean scores varied from 28 per cent to 72 per cent, the central tendency being near to 50 per cent.) Since comparisons have always been made upon the *same* pupils (or presumably equal groups of pupils), and the variables in our interpretations have been (*a*) differences in the marks of different scorers, (*b*) differences in the marks on consecutive examinations scored by the same person, or (*c*) comparisons of subjective and objective forms of the same questions and examinations, the matter of departures from the regular technique of the Regents' examinations, although admittedly great, would not seem to be an important factor in our interpretations unless present conclusions should be directed at criticisms of the Regents' examinations, per se. This intention is disclaimed. The Regents' examinations as

here used are merely a means to an end, not the point of attack.

However, to return to the question of the competency of the teachers who graded the twenty-six sets of papers represented in Tables V and VII, all of the teachers were college-trained, with graduate work in the social studies and with considerable school experience. They were selected from a lot of 100 teachers in a course on objective examination methods upon the basis of their relative fitness. Certainly these were at least an average group of public-school teachers, and very probably they may be regarded as a rather superior group of teachers.

A second and very important consideration in the discussion of the ability of the twenty-six markers was the fact that these teachers were just completing a six-weeks' course of study in which especial attention had been paid to the sources of error in the grading of examination papers. All of the principal studies of the fallibilities of teachers' marks had been discussed at length, and many suggestions for minimizing such unreliability had been considered. The effects of such studies are undoubtedly present in the markings of the subjective examinations of the present investigation and unquestionably would have tended to reduce error rather than increase the unreliabilities reported here.

By and large, there is considerable warrant in assuming that the reliability coefficients reported for the subjective examinations are conservative statements to the truth of the matter.

Discussion of Table VIII. Table VIII may be summarized by four statements:

(1) The fundamental fact brought out by Table VIII is that the objective examinations required somewhat less time than the subjective forms of the same sets of questions. The time ratios based upon a comparison of the actual

times needed to answer 1424 subjective examinations and 2342 objective examinations are as follows:

$$\frac{\text{Average Working-Times: Subjective Examinations}}{\text{Average Working-Times: Objective Examinations}} = \frac{37.0}{31.4} = 1.18$$

The reversed ratio is .85.

(2) Reliability per *unit of working-time* is a fairer method of comparing reliabilities than reliability per ten-question examinations.

If we substitute the average reliability coefficient of the objective examinations (.65) in the Spearman-Brown prophecy formula,

$$r_{nn} = \frac{nr}{1+(n-1)r}$$

letting n be equal to 1.18, we find the reliability of the objective examinations for the same working-time as the average subjective examination to be .69. The value .69 is from many points of view a more valid figure for comparison with the corresponding reliabilities of the subjective examinations (.43 or .40 of Table V; the choice being uncertain) since working-times are then constant.

(3) Attention is again directed to the fact that the objective examinations of the present study were limited to the same theory and method of sampling involved in the subjective forms of the same examination. Could the "extensive" theory of sampling have been used, the resulting reliabilities per unit of working-time probably would have been very much higher. (See Chapter IV.)

(4) The fact that the average working-times were but 37 and 31 minutes, respectively, for the subjective and objective examinations may raise the question whether the pupils were given sufficient time for answering. The Regents' examinations allow 180 minutes working-time. In answer to this objection it can be stated that the pupils were given as much time as they wished. *Also, all unfinished papers were discarded.* Inspection of the papers shows that the

pupils had apparently written all that they knew about each question and that there was no manifest evidence of undue haste or lack of effort. Since these pupils had not studied the regular program of social subjects of New York high schools, it is to be expected that they should make relatively lower marks and use less time in answering. Actually, there were many pupils who wrote for far longer times than the mean times reported. This is true of both objective and subjective types.

Discussion of Table IX. Table IX may be summarized by the following statements:

(1) Table IX shows that the two types of examinations were equal in economy of time needed for correction, the average correction times being 4.2 (subjective) and 4.1 (objective).

(2) The only point of relative superiority of the one over the other lies in the fact that cheap clerical help could have been (and was in part) employed for reading the objective examinations, but the reading of the subjective examinations required more costly "expert" readers.

(3) The scoring of the present objective examinations was far more laborious than would have been the case had the examinations been built de novo rather than remodelled from subjective forms. The present correction times are far from typical of objective examinations. In fact, stated per score-unit, they are from two to four times longer than the types of objective examinations described in Chapter IV, where a uniform mechanical arrangement of test items was employed. Definite experimental evidence was obtained on this point, but was omitted here because it was considered to be of minor importance.

V. SUMMARY AND CONCLUSIONS

(1) Thirteen lots of fifty examinations of the New York Regents for the year 1923 were read independently by two

teachers. The mean r between the marks of the two teachers was .74. Thirteen lots of the 1924 examinations were treated in the same manner, the mean r proving to be .69. The mean r of the twenty-six lots was .72. These r's may be called the reliability coefficients of scoring, or coefficients of subjectivity (columns (*a*) and (*b*) of Table V).

(2) Twenty-six lots of fifty examinations where the same pupils took both the 1923 and 1924 editions of the New York Regents' examinations were read by the same teacher. The mean r was found to be .43. These coefficients show the combined effects of subjectivity and errors due to the limited sampling possible in eight- and ten-question examinations (columns (*c*) and (*d*) of Table V).

(3) Twenty-six lots of examination papers where the 1923 and 1924 examinations were taken by the same pupils, but the papers for the two different years were read by different teachers. The mean r was .40 (columns (*e*) and (*f*) of Table V).

(4) Reliability coefficients of .72, .43, and .40 represent alienation from perfect reliability of 69 per cent, 90 per cent, and 92 per cent respectively; i.e., the standard errors of estimate are reduced over the situation of chance assignment of marks by 31 per cent, 10 per cent, and 8 per cent, respectively. (See Appendix.)

(5) The factor of subjectivity is completely reducible if objective forms of questioning are employed.

(6) Eight New York Regents' examinations turned into objective form showed higher reliability than the original form, viz., a mean r of .65 in comparison with .43 and .40 for the comparable situations for the original editions (Table VI).

(7) Twenty-six lots of examination papers where two teachers read each paper independently showed an average difference of mean scores for the examinations of 6.0 points.

At least 10 per cent of the differences in the mean marks assigned by two different teachers are equal to one or more standard deviations. In other words, the two distributions, if superimposed, would show a net displacement of 1.00 S. D. or more in 3 sets of examinations from the lot of twenty-six (Table VII).

(8) The Regents' examinations in original form appear to be markedly unequal in difficulty from year to year (Table VII).

(9) The variability of the marks assigned by two independent readers of the same papers was found to be very different in many cases, differences as great as 8.7 points being observed. The mean difference was about 3 score-points (Table VII).

(10) The mean working-time for 1424 subjective examinations was found to be 37.0 minutes. The mean working-time for the objective forms of the same examinations proved to be 31.4 minutes for 2342 papers. This is equivalent to saying that the pupils answered the objective examinations 1.18 times as fast, or that the objective examinations could have been made 1.18 as long as the subjective without increasing the length of the examination in working time (Table VIII).

(11) The times needed for scoring the papers were equal for objective and subjective variates of the same examinations, the mean scoring time being slightly more than 4 minutes in each case (Table IX).

(12) If the objective and subjective examinations are compared for reliability per equal working-times, the r's would be about .69 and .40 to .43, respectively.

(13) These results are not to be taken as criticisms of the New York Regents' examinations, per se, for many reasons:

(*a*) The pupils had not studied the New York course of study.

(*b*) No direct evidence of the relative validities of subjective and objective forms is available.

(*c*) The readers of the New York Regents' examinations as used in the present study were not the official readers, and the administration of the examinations was not regular.

(14) The Regents' examinations were employed merely as typical examinations of the traditional type in order to study the possibility of improving such examinations through the adoption of the recent proposals for objective forms of questions. The evaluation of these examinations was indirect, impersonal, and incidental rather than an attempt to discredit such examinations.

(15) The present data on the objective examinations actually employed are not in close harmony with the results reported by Wood, Toops, Ruch and Stoddard, DeGraff, *et al*, since the objective examinations were forced into many limiting conditions on account of the fact that they were "translations" of the official Regents' examinations rather than de novo attempts at objective-examination construction. In particular, the method and theory of sampling as related to good examination practice were not open to choice, the "translation" of fixed examinations into objective form forcing the use of a theory of sampling believed by the writer to be inferior. A scheme for sampling in examinations was presented and discussed with the recommendation that "extensive" sampling be employed rather than the "intensive" sampling of the Regents' examinations.

CHAPTER IV

STUDIES ON THE RELATIVE MERITS OF RECALL, MULTIPLE-RESPONSE, AND TRUE-FALSE TECHNIQUES WITH SPECIAL REFERENCE TO CORRECTIONS FOR CHANCE[1]

I. INTRODUCTION

Various types of objective examinations. The rapid growth of the objective examination in education has developed several types of tests, chief among which are the true-false, multiple-choice, matching-exercise, and the single- and multiple-completion. A special form of the completion test is the simple recall, where the statement is so made that a short answer, a word, or a phrase is needed to complete the sentence. This omitted word or phrase is essentially the key-word of the sentence and represents the item of information the examiner wishes to secure from the one examined.

Each of the first three types mentioned involves the element of chance or guessing, since from two to seven items are presented for choice and must of necessity introduce the element of suggestion of the correct answer. At the point where certain knowledge ends, choice or guessing begins, and it has been the determination of this point that has been the bone of contention among test workers, educators, and psychologists. Reasoning from the laws of chance, it has been determined that the most logical method of correction should be by means of the formula, $R-\frac{W}{n-1}$, where R represents the number of items answered correctly, W the number of incorrect answers, and n the number of choices offered. This method as applied to the ordinary tests of from thirty to one

[1]An abstract of a thesis presented to the Graduate College of the State University of Iowa in partial fulfillment of the requirements for the degree of Doctor of Philosophy by M. H. DeGraff, 1925, under the title, *The Validity of Corrections for Chance in Objective Examinations Involving Multiple-Choice.*

hundred items has been attacked on both logical and experimental grounds and has likewise been supported on the same grounds. The chief difficulties are the location of that point of certain knowledge and the method of correction for guessing beyond that point.

Another question arising out of the now general use of the objective test is the type of test to use. The true-false is probably used most frequently, since it is comparatively easy to prepare, although this opinion might be brought under question if all the requirements of a good test were considered carefully. Again, it involves little difficulty in administering. On the other hand, the true-false test has been the one most often attacked on logical, pedagogical, and psychological grounds, as well as upon the basis of its unreliability. Where the multiple-choice or multiple-response types of test have been used, a question has arisen regarding the number of choices to be offered in order to test most fairly the knowledge of the student. Is the two-, the five-, or the seven-response test the most reliable measure of the information of the one tested? What is the relation of the time required to the type of test which might be used? Which one can be scored with the least expenditure of time and energy? These are some of the questions which present themselves to the administrator, supervisor, teacher, or research department as they formulate their objective-testing program.

It is the purpose of this study to attack the problem experimentally with a view to solving the following questions:

(1) What is the validity of the correction formula $R-\frac{W}{n-1}$ when applied to an actual school subject?

(2) Which of the five following types of tests are the more reliable techniques of measurement: true-false, two-response, three-response, five-response, or seven-response?

(3) What is the effect upon reliability of careful and explicit directions, e. g., (*a*) to guess, (*b*) not to guess?

(4) What are the relative difficulties of the recall, the true-false, and the recognition types?

(5) How much time is required for attempting to answer one hundred items of the recall type in comparison with one hundred items, respectively, of the five types mentioned above; i. e., how many items in the true-false form can be answered per unit of time for a recall item?

(6) What are the relative places of the six types so far as time-consumption in scoring is concerned?

Related studies. It is not the purpose of this study to report here a critical analysis of the somewhat numerous studies which have been made in the field as outlined in the preceding chapter. The conclusions reached in a few of the most closely allied investigations will be discussed briefly.

Brinkley[1] found that the types ranged in order of decreasing difficulty as follows: (1) completion, (2) true-false, (3) multiple-choice.

The true-false gave the most valid results when corrected by $R-W$ rather than merely "rights" or $R-\frac{1}{2}W$. (There would seem to be no valid argument for the method $R-\frac{1}{2}W$ as used by Brinkley.) The three-response was the most valid under the $R-\frac{1}{2}W$ correction, while with the four-response and five-response, "score = rights" gave the most valid result. He criticises the true-false on the ground of guessing and its attendant injustice and quotes Chapman and West in support of his criticism. He reports the time for the various tests as follows:

Test	Length in Minutes (Average for Group)
True-False	.31
Multiple-Choice	.48
Completion	.46
Word-Phrase Answer	.31
Arrangement	.22
Essay	.72

[1] S. G. Brinkley, "Values of New-Type Examinations in the High School." *Teachers College Contributions to Education*, No. 161, 1924.

West[1] found that the $R-W$ method of scoring unduly penalized the subjects as a whole. No criterion of guessing was used in West's study except the indication by the subjects as to when they guessed and when they were certain. However, in the opinion of the present writer the inability or unwillingness of the subjects to say when they were actually guessing might be a contributive factor in the seeming injustice. The experiment was conducted with 174 cases on a short (50-item) synonym and antonym test.

Hahn,[2] on the basis of twenty-five complete drawings of twenty-five black and twenty-five white buttons out of a bag containing thirty-five white and thirty-five black buttons and the subsequent scoring of each of the twenty-five trials according to the $R-W$ method, makes this statement:

> What does the final score of such tests (true-false) represent? No one knows. That it cannot even approximately represent real ability or actual achievement has been shown.

Barthelmess,[3] in replying to the criticism brought forward by Hahn, sums up the situation as follows:

> The best test of any test is the correlation with a criterion. If this correlation is satisfactory, we can forget all minor criticisms concerning chance, etc.

The calculations of the probabilities of correct answers ranging from zero to perfect lead Asker[4] to state:

> But when part of the test is known and the rest is answered through guessing, there are rather large chances for undeserving individuals to make a passing grade and for deserving pupils to fail. Especially is this true of a test with only two possible answers.

In Chapman's study[5], the injustice done by the right-

[1] P. V. West, "Critical Study of the R – W Method of Correcting True-False," *Journal of Educational Research*, Vol. 8, pp. 1-9.
[2] H. H. Hahn, "A Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 6, pp. 235-240.
[3] H. M. Barthelmess, "Reply to a Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 6, pp. 355-359.
[4] Wm. Asker, "The Reliability of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 9, pp. 234-240.
[5] J. C. Chapman, "Individual Justice and Guessing in the True-False Examination," *Journal of Applied Psychology*, Vol. 6, pp. 342-349.

minus-wrong method is illustrated in the case of one subject's taking the test of fifty items, all of which were supposed to be answered. When the subject had completed fourteen items, the examiner was informed that this was the extent of the known answers. Of the fourteen, thirteen were correct, giving a true score of twelve, but when the subject completed the test, the score resulting from the correction by $R-W$ was zero. Chapman goes on with a hypothetical situation of tests of thirty, sixty, and ninety items to show the inoperative effect of chance. Even when directed not to guess, the subject with the sporting disposition will invariably do a certain amount of guessing, and Chapman concludes that "by making the penalty for errors sufficiently great, we can probably deter the wise, but we shall lower the reliability."

In the study made by Gates,[1] where the true-false test, the essay examination, written papers, oral quizzes, special conferences, etc., were examined, it was found that the true-false test correlated highest with the criterion (the sum of all the measures arbitrarily weighted), and the following conclusion was made: "The true-false test thus appears, all things considered, to be the most reliable single measure of achievement." The entrance of the factor of spurious self-correlation of a part with the whole would seem to weaken somewhat the conclusion.

Batson[2] compared true-false tests totalling two hundred items with essay examinations totalling thirty-one questions and found that on the basis of a population of 113 the two show fairly close agreement when the results were weighted or distributed according to the five-point system of grading.

In the experiment by Remmers,[3] four groups were divided equally as to ability as shown by the Otis Group Test; two

---

[1]A. I. Gates, "The True-False Test as a Measure of Achievement in College Courses," *Journal of Educational Psychology*, Vol. 12, pp. 276 ff.

[2]W. H. Batson, "The Reliability of the True-False Form of Examination," *Educational Administration and Supervision*, Vol. 10, pp. 95-102.

[3]H. H. Remmers, "Relative Difficulty of True-False, Multiple Choice, and Completion Tests," *Journal of Educational Psychology*, Vol. 14, pp. 366-371.

groups being given multiple-choice (three-response), one group, true-false, and one, the single-completion type. The difficulty as shown by the average scores placed the true-false as the hardest, the multiple-choice easiest, and the single-completion between these two.

Toops,[1] in his study of trade tests in education, took occasion to compare the time required as well as the reliability of three types of objective tests, namely, the recall, the five-response, and the true-false. This study was made with college classes (one hundred twenty-four cases), and hence the range of talent was rather limited, thus affecting both the reliabilities of the tests used and the time required. A comparison of the results found by Toops and those of the present study will be presented in a later section of this chapter.

The study made by Ruch and Stoddard[2] closely parallels this present study, and, in fact, was the basis for suggesting it. The first experiment was carried out on a comparatively small number of cases, 562, included only four types besides the recall, and did not attempt to check up the influence of directions in regard to guessing as against non-guessing. These facts, in addition to the erratic behavior of one or two of the tests, made advisable the more careful and extended study of the problem. As in the case of the Toops study, the results obtained in the Ruch-Stoddard experiment will be compared with the present study in a later section of this chapter.

II. METHOD OF PROCEDURE

Selection of materials. The field of United States history was chosen on the ground that this subject is taught in the seventh, eighth, and either eleventh or twelfth grades. Thus the study could be staged at the various levels, insur-

[1]H. A. Toops, "Trade Tests in Education," *Teachers College Contributions to Education*, No. 115 (1921).

[2]G. M. Ruch and G. D. Stoddard, "Comparative Reliabilities of Five Types of Objective Examinations," *Journal of Educational Psychology*, February, 1925.

ing a sufficient range of talent, in addition to a large number of cases. The material taught at the various levels is at least fairly comparable and consecutive, enabling the results to be pooled without treatment of scores. In addition, history material, when taken on the basis of general information, lends itself very well to objective examinations of the types desired for this study.

In order to secure the material for the statements to be used in the tests, three sources were tapped: (*a*) certain statements in the matching tests, which had been constructed by co-workers in the present Commonwealth investigations, (*b*) questions asked by the New York Board of Regents, and (*c*) three history textbooks which have widespread use either in junior or senior high schools. The three textbooks were: Muzzey's *An American History*, Gordy's *History of the United States*, and Beard and Bagley's *History of the American People*.

Certain definite requirements were set up as to the suitability of statements to be used. These requirements were as follows:

(1) The incompleted statement should have objectivity; that is, there should be one or at least a very limited number of correct answers possible.

(2) The statement should not be ambiguous as to the kind of answer required, that is, as to whether a name, an event, or a date was indicated.

(3) The tests should admit of easy and rapid scoring.

(4) The nature of the statement and answer should be such that seven possible and plausible answers would be available, in order that the seven-response test might be constructed on the same material.

(5) The range of difficulty should be great enough to include questions easy enough for pupils of the average seventh-grade training, and some should be difficult enough to preclude perfect scores even by the best high-school seniors.

(6) The range of subject-matter should be great enough to encompass the work being done at all four grade-levels.

(7) The range of the type of questions should include statements to be answered by dates, events, names of men, names of states, cities, countries, policies, acts of legislation, slogans, etc.

It might be mentioned in passing that of these seven requirements the one of availability of seven possible and *plausible* answers was by far the most difficult single task in the construction of the test material.

Before the seven responses were selected, 250 questions were typed on three-by-five cards and submitted to six judges, who were to rate each question on two scores: first, the desirability of the item as a test question; and second, the relative difficulty of the question, dividing the entire list into ten classes of difficulty. These six judgments were made independently, and in each case the double score was written as a fraction on the back of the card. For example, an item judged as one of the most difficult and also most desirable as a test item would be marked 10/1, while an item considered as very easy and not suitable for testing would be marked 1/4. These six judgments were averaged, thus dividing the items into ten approximately equal groups as to difficulty. Those marked three or four as to desirability were discarded.

This procedure resulted finally in two hundred items arranged as nearly in the order of difficulty as could be done without submitting them to the process of trial testing for difficulty. Inasmuch as the tests were planned to act as a medium for statistical analysis of the validity of the correction for chance and of comparing the various types, it was not deemed necessary to resort to the preliminary-testing process. This is not done by the teacher or supervisor as she constructs her objective examinations, and it seemed wiser to parallel such conditions rather than the conditions of the standardized test.

Having secured the two hundred most desirable items, the next step was to divide these into two groups of one hundred items each of approximately equal difficulty. Since they had been graded for difficulty by the method of pooled judgments, the method of odds and evens would have resulted in a slightly higher summation of difficulty for the evens than for the odds. The items were therefore divided as follows: Form A—items 1, 4, 5, 8, 9, 12, 13, etc.; Form B—items 2, 3, 6, 7, 10, 11, etc. That this method gave two forms of equal difficulty is evidenced by the mean scores of 27.8 and 27.6, respectively.

The construction of the recognition and true-false types was the next step in the problem. The seven-response test was constructed first in order that the number of choices might later be reduced to forms with fewer choices, rather than building up from the two-response to the seven-response. The responses were selected with extreme care in order that all wrong alternatives might present as nearly equal plausibility as the correct answers, assuming ignorance on the part of the pupil. In order to secure this equality, answers were chosen which represented a close relationship in point of time, association, or likeness of sound or purpose. For example, the statement "The first Pilgrims were brought to American shores in the ship named .....?" has for its possible answers "Santa Maria," which carries the association of Columbus coming to American shores; "Trent," which is learned in the Mason-Slidell or Trent affair; "Mayflower," the correct answer; "Half Moon," associated with the voyage and landing of Henry Hudson; "Monitor," "Maine," and "Lusitania," as being names of ships whose names are or should be prominent in the memory of the pupil.

That some of the responses are related to or connected with the question only indistinctly cannot be denied, but is it not true that this situation actually paralells the knowledge condition of the pupil? It would be impossible to

secure two or more answers that would be exactly equal in probability if the correct answer were definitely determined. As we increase the number of choices, we must of necessity move farther our from the center where absolute certainty lies. However, the attempt was made to keep the choices as close together in this particular as possible.

To formulate the five-response type, two of the most distantly connected answers were dropped. Two more were eliminated to form the three-response, and still another to get the two-response forms. By using the correct answers of the two-response for one-half of the items and the incorrect answers for the other half, the true-false forms were obtained. Thus each of the six different test types was made up of exactly the same items, the number of alternative responses forming the only change. A sample question, used in all six types, follows:

Recall: Eli Whitney is noted for his invention of the.....

7-response: Eli Whitney is noted for his invention of the (1) steamboat (2) spinning jenny (3) cotton gin (4) telegraph (5) telephone (6) printing press (7) steam engine.

5-response: Eli Whitney is noted for his invention of the (1) steamboat (2) spinning jenny (3) cotton gin (4) telegraph (5) telephone.

3-response: Eli Whitney is noted for his invention of the (1) spinning jenny (2) cotton gin (3) telegraph.

2-response: Eli Whitney is noted for his invention of the (1) spinning jenny (2) cotton gin.

True-false: Eli Whitney is noted for his invention of the spinning jenny.

The typographical construction of the tests may be of interest to the reader. The type was eleven point on thirteen, giving space between the lines of the question, while a single-pica space was inserted between questions to prevent confusion of two questions. The questions were so phrased that the omission came at the end of the statement in every case. This was done in order to admit of all answers being placed in the right-hand margin of the sheet, the idea being

that this would render the answering easier for the subject and also facilitate the scoring.

Since it was a part of the study to determine the effect of directions in regard to guessing, one-half the tests were prepared with directions to guess and one-half with directions to answer only those of which the students were reasonably certain. A copy of these directions is shown here.

SERIES IV TEST 15

GENERAL INFORMATION TEST IN AMERICAN HISTORY
RECOGNITION: 7-RESPONSE

For experimental purposes only
Commonwealth Investigation of Examinations in the Social Sciences
University of Iowa

Name.....Age.....Grade.....

Sex.....School.....
(City) (State)

Date of this Test.....Date of Birth.....
(Month) (Day) (Year)

Directions:

On the following pages there are two sets of 100 questions about American history.

Seven possible answers are given to each question. You are to select the best answer and then write the *number* of the answer (not the answer itself) on the dotted line at the right of each question, as shown in the following samples:

1 The United States is in (1) South America (2) Asia (3) Africa (4) Europe (5) North America (6) Australia (7) Central America ..5..

2 Benjamin Franklin was an (1) Englishman (2) Frenchman (3) Irishman (4) African (5) German (6) Chinaman (7) American . . . .

What number belongs on the blank for the second sample?
Write "7" on the blank now.

☞NOTE CAREFULLY: Do not leave any question unanswered. If you don't know, guess. It is better to guess than to leave a question blank because you have one chance in seven of getting it right by pure guessing. You should try to make as logical or shrewd a guess as possible.

When you finish, record the time taken in the space provided at the end of the examination.

REMEMBER: Try to answer every question. Guess if you do not know.

Wait for the signal to begin!

The companion test (Test 16, Series IV) was identical, except for those portions of the directions set in bold-faced type. These were replaced by the following:

☞NOTE CAREFULLY: If you are in doubt about the answer to any question, leave it blank. Do not guess! You will be penalized for all wrong answers. The tests are scored in such a way that you will lose more than you will gain by guessing.

When you finish, record the time taken in the space provided at the end of the examination.

REMEMBER: Do not guess. Answer only those that you are reasonably sure about.

Wait for the signal to begin!

In general, these instructions were followed explicitly. The writer, in administering part of the tests, observed no hesitancy on the part of those taking the tests, and in very few of the three thousand papers corrected were there evidences of confusion or misreading.

The experimental procedure. The conditions of the experiment were as follows:

First sitting: Recall Test, Form A, was administered to all pupils.

Second sitting: On the same or following day, Recall Test, Form B, was administered to all pupils.

Third sitting: Some *one* of the ten recognition tests was administered to each pupil. In order to secure random

sampling on the various types, the tests were sent to the coöperating schools arranged in such manner that the ten different forms of test booklets ran in cycles, repeating the cycle every tenth booklet, thus, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, etc. The actual numbering of the tests was from 15 to 24. The test administrators were instructed to "deal off the top of the pile," thus insuring that in a class of fifty, ten pupils would be taking the seven-response, ten the five-response, ten the three-response, ten the two-response, and ten the true-false, and, further, that each group would be divided equally between those working under instructions to guess and those working under instructions not to guess. In summary, a class of fifty pupils would thus be divided *at random* into ten sub-groups of five each.

It was the purpose of the writer to secure about thirty-five hundred cases paired completely through the three testings, and for this reason complete sets for approximately fifty-five hundred pupils were sent out to the coöperating schools. However, because of the expected mortality on a testing program of three days and of unfortunate misunderstandings on the part of the coöperators, the number of cases was reduced to twenty-four hundred and fifty-three. The numbers in the ten sub-groups ranged from 229 to 281 subjects.

The tests were administered in seventh, eighth, eleventh, and twelfth grades in school systems in Minnesota, Illinois, Iowa, Missouri, Oklahoma, Texas, Arizona, and California. The spread of grades and localities insured a wide range of talent and is doubtless a factor in the higher correlations obtained in this study than those of Toops or of Ruch and Stoddard. Very careful and detailed instructions were mimeographed and sent to those administering the tests in order that similar conditions of testing might obtain in all places. The matter of recording of time is especially emphasized in the directions in order that determinations of the number of items per unit of working-time might be made as accurate as possible.

The statistical treatment of the results may be outlined in brief as follows:

(1) A comparison of the means and sigmas of the various types to determine the relative difficulty of each.

(2) Using the recall tests, forms A and B, as the control test or criterion to correlate the respective recognition forms against the criteria, as well as to determine the coefficient of reliability between the two forms of each type.

(3) To repeat the above correlations when the scores had been corrected for chance.

(4) To correct the correlations for attenuation due to errors of measurement.

(5) To correct the correlations of the recognition tests by means of the Spearman-Brown formula, $r_{nn}=\frac{nr}{1+(n-1)r}$, for the number of items of the recognition types per unit of time for one hundred recall items.

(6) To present percentiles of time required for the six types of tests used.

(7) To present the time of the average scorer per one hundred items of the various types.

III. RESULTS OF THE EXPERIMENT

Number of cases. The number of cases paired throughout the three testings was 2453. These are divided into ten sub-groups as follows: five groups under instructions to guess, working on the seven-, five-, three-, two-response, and true-false types, and five groups under instructions not to guess on the five tests respectively.

TABLE XI

NUMBERS OF PUPILS USED IN THE EXPERIMENT

DIRECTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	233	236	246	223	239
Do Not Guess	229	274	229	281	263

The criterion. As stated, the criterion or control factor was the recall test in two forms. When the scores of the 2453 cases were pooled, the correlation of the two forms was .95 with a P. E. of .0013, indicating that the test is a comparatively reliable one. A comparison of the mean scores (27.6 and 27.8 respectively) indicates that Form A and Form B are very nearly equal in difficulty.

Influence of corrections for chance. Tables XII to XXIII show the comparisons and the direction of the differences between correlations of corrected scores and correlations of uncorrected scores, as well as the differences between results under instructions to guess and not to guess. These are given separately for the correlations between Recall A and Recognition A, Recall B and Recognition B, and Recognition A and Recognition B, in order that more intimate comparisons may be made easily.

Tables XII to XVII present comparisons of results of correction and non-correction, while Tables XVIII to XXIII compare the effect of instructions to guess at all items, in contrast to instructions to answer only those items of which the subject is reasonably certain. The differences in all cases are denoted by the plus sign if the r of the *corrected* scores or of the scores made under directions *not to guess* is the larger and by the minus sign in all other cases.

Of the ten comparisons of Tables XII and XIII only one shows an advantage in favor of scoring without correction, and this is less than one-fifth of the probable error of either correlation. The other nine show advantages in favor of using the correction for chance, and the amounts of difference vary from one-third to six times the probable error of either correlation. The greatest amount of difference is shown in the case of the true-false under instructions not to guess and corrected for chance.

TABLE XII

COMPARISON OF CORRELATIONS: "GUESS" INSTRUCTIONS, RECALL A VS. RECOGNITION A

	Responses				
	7	5	3	2	True-False
Uncorrected for Chance	.871	.907	.838	.859	.804
Corrected	.873	.910	.848	.865	.839
Difference[1]	+.002	+.003	+.05	+.006	+.035

[1]The plus sign denotes superiority of correlation of corrected scores over correlation of uncorrected scores. The minus sign indicates the reverse situation. This practice is uniform for Tables XII to XVII.

TABLE XIII

COMPARISON OF CORRELATIONS: "DO NOT GUESS" INSTRUCTIONS, RECALL A VS. RECOGNITION A

	Responses				
	7	5	3	2	True-False
Uncorrected for Chance	.927	.891	.845	.740	.749
Corrected	.926	.918	.915	.775	.860
Difference	−.001	+.027	+.07	+.035	+.111

The comparison of Recognition Form B correlated with the criterion, Recall B, (Tables XIV and XV) shows approximately the same, except that in no case is the correlation of the uncorrected scores higher than that of the corrected. The true-false type again shows the largest amount of advantage, but in this case it occurs in the scores of those subjects instructed to guess. The amount of difference in the case of the two-response, "do not guess," is only slightly less than that of the true-false, "guess." In no case is the difference less than three times the probable error of the smaller of the correlations.

TABLE XIV

COMPARISON OF CORRELATIONS: "GUESS" INSTRUCTIONS, RECALL B VS. RECOGNITION B

	RESPONSES				
	7	5	3	2	True-False
Uncorrected for Chance	.816	.860	.797	.735	.675
Corrected	.861	.903	.875	.806	.801
Difference	+.045	+.043	+.078	+.071	+.126

TABLE XV

COMPARISON OF CORRELATIONS: "DO NOT GUESS" INSTRUCTIONS, RECALL B VS. RECOGNITION B

	RESPONSES				
	7	5	3	2	True-False
Uncorrected for Chance	.872	.836	.852	.752	.768
Corrected	.898	.870	.902	.868	.856
Difference	+.026	+.034	+.050	+.116	+.088

Reliability coefficients of the tests. The reliability coefficients of the recognition types are presented in Tables XVI and XVII. In the case of both the two-response and

the true-false, "do not guess," the correlation of the uncorrected scores is slightly higher than that of the corrected scores, but is less than three times the probable error in the case of the true-false and just equal to the probable error in the case of the two-response. All other differences are in favor of correction for chance, although none is larger than three times the probable error.

TABLE XVI

COMPARISON OF CORRELATIONS: "GUESS" INSTRUCTIONS, RECOGNITION A VS. RECOGNITION B

	RESPONSES				
	7	5	3	2	True-False
Uncorrected for Chance	.800	.864	.837	.745	.641
Corrected	.839	.902	.858	.864	.780
Difference	+.039	+.038	+.021	+.019	+.139

TABLE XVII

COMPARISON OF CORRELATIONS: "DO NOT GUESS" INSTRUCTIONS, RECOGNITION A VS. RECOGNITION B

	RESPONSES				
	7	5	3	2	True-False
Uncorrected for Chance	.886	.862	.886	.859	.885
Corrected	.907	.882	.890	.843	.837
Difference	+.021	+.020	+.004	−.016	−.048

Instructions relative to guessing. In comparing the correlations of scores made when the instructions are to guess with those when the subjects are directed not to guess, there is much more variance than in the case of correction versus non-correction. The differences favor slightly "do not guess" when corrections for chance are not used (Tables XVIII and XX) but indicate considerable advantage for "do not guess" instructions when corrections are applied (Tables XIX and XXI).

The reliability coefficients of Tables XXII and XXIII would seem to indicate real superiority of "do not guess" directions where no corrections are applied to 2-response and true-false tests.

TABLE XVIII

COMPARISON OF CORRELATIONS: UNCORRECTED FOR CHANCE. RECALL A VS. RECOGNITION A

INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.871	.907	.838	.859	.804
Do not guess	.927	.891	.845	.740	.749
Difference[1]	+.056	−.016	+.007	−.119	−.055

[1]Plus sign denotes superiority of correlation of scores made under "do not guess" over the correlation of scores made under "guess" instructions. The minus sign denotes the opposite. This practice is uniform for Tables XVIII to XXIII.

TABLE XIX

COMPARISON OF CORRELATIONS: CORRECTED FOR CHANCE. RECALL A VS. RECOGNITION A

INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.873	.910	.848	.865	.839
Do not guess	.926	.918	.915	.775	.860
Difference	+.053	+.008	+.067	+.090	+.021

TABLE XX

COMPARISON OF CORRELATIONS: UNCORRECTED FOR CHANCE. RECALL B VS. RECOGNITION B

INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.816	.860	.797	.735	.675
Do not guess	.872	.836	.852	.752	.768
Difference	+.056	−.024	+.055	+.017	+.083

TABLE XXI

COMPARISON OF CORRELATIONS: CORRECTED FOR CHANCE.
RECALL B VS. RECOGNITION B

INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.861	.903	.875	.806	.801
Do not guess	.898	.870	.902	.868	.856
Difference	+.037	−.033	+.027	+.062	+.055

TABLE XXII

COMPARISON OF CORRELATIONS: UNCORRECTED FOR CHANCE.
RECOGNITION A VS. RECOGNITION B

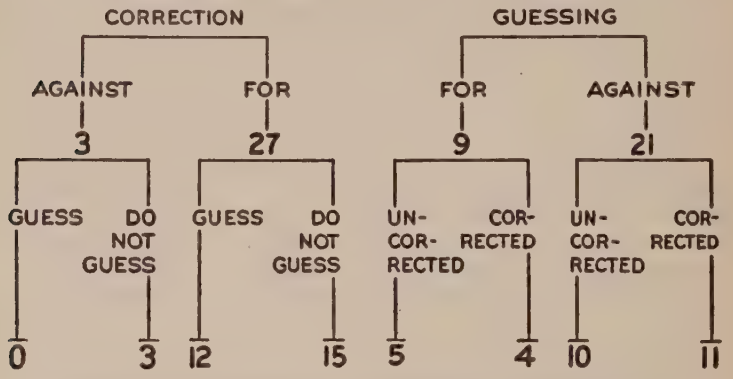
INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.800	.864	.837	.745	.641
Do not guess	.886	.862	.886	.859	.884
Difference	+.086	−.002	+.049	+.114	+.243

TABLE XXIII

COMPARISON OF CORRELATIONS: CORRECTED FOR CHANCE.
RECOGNITION A VS. RECOGNITION B

INSTRUCTIONS	RESPONSES				
	7	5	3	2	True-False
Guess	.839	.902	.858	.864	.780
Do not guess	.907	.882	.890	.843	.837
Difference	+.068	−.020	+.032	−.021	−.043

The figure below is to be read as follows: Of the thirty comparisons of correlations of corrected and uncorrected scores, three showed the latter to be larger. All of these were under instruction not to guess. Twenty-seven cases revealed the superiority of the corrected over the uncorrected scores,



SCHEMATIC COMPARISON OF DIFFERENCES (TABLES XII-XXIII)

and of these, fifteen were of groups instructed not to guess. When comparisons are made between guessing and not guessing, it is found that the correlations of scores made when subjects do not guess are higher in twenty-one cases and lower in but nine cases than when the pupils guess. This summary supports the conclusion that it is better to use correction for chance and instruct not to guess.

Of the sixty differences noted in Tables XII to XXIII, twenty-six are in favor of "do not guess" and correction for chance; twenty-two in favor of "guess" and correction for chance; five in favor of "guess" and non-correction; and seven in favor of "do not guess" and non-correction.

Intercorrelations of the tests. In order to facilitate comparison of the correlations of all types of the examinations under both sets of instructions and with scores corrected and uncorrected for chance, all coefficients of correlation (raw) with their P. E.'s are combined in Table XXIV.

TABLE XXIV

INTERCORRELATIONS, CORRECTED AND UNCORRECTED FOR CHANCE, FOR ALL TEN TESTS USED

TYPE OF TEST	RECALL A VS. RECOGNITION A		RECALL B VS. RECOGNITION B		RECOGNITION A VS. RECOGNITION B	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
7-response (g)[1]	.871 ± .011	.873 ± .011	.816 ± .015	.861 ± .111	.800 ± .016	.839 ± .013
7-response (n)[2]	.927 ± .006	.926 ± .006	.872 ± .012	.898 ± .009	.886 ± .010	.907 ± .008
5-response (g)	.907 ± .008	.910 ± .008	.860 ± .011	.903 ± .008	.864 ± .011	.902 ± .008
5-response (n)	.891 ± .009	.918 ± .007	.836 ± .013	.870 ± .010	.862 ± .011	.882 ± .010
3-response (g)	.838 ± .013	.848 ± .012	.797 ± .016	.875 ± .010	.837 ± .013	.858 ± .011
3-response (n)	.845 ± .014	.915 ± .007	.852 ± .012	.902 ± .008	.886 ± .010	.890 ± .009
2-response (g)	.859 ± .012	.865 ± .011	.735 ± .021	.806 ± .016	.745 ± .020	.864 ± .011
2-response (n)	.740 ± .018	.775 ± .016	.752 ± .018	.868 ± .010	.859 ± .011	.843 ± .012
True-false (g)	.804 ± .015	.839 ± .013	.675 ± .024	.801 ± .016	.641 ± .026	.780 ± .017
True-false (n)	.749 ± .018	.860 ± .011	.768 ± .017	.856 ± .011	.884 ± .009	.837 ± .013

Correlation of Recall A vs. Recall B ..... .950 ± .001
Coefficient of reliability ..... .974 ± .001

[1](g) indicates the tests taken under instructions to guess.
[2](n) indicates the tests taken under instructions not to guess.

Corrections for attenuation. In order to eliminate the chance errors, thus ruling out the factor of unreliability, the correlations of the recall and the recognition types were corrected for attenuation by means of Spearman's shorter formula:

$$r_{\infty\infty} = \frac{\sqrt{r_{x_1y_1} \cdot r_{x_2y_2}}}{\sqrt{r_{x_1x_2} \cdot r_{y_1y_2}}}$$

where $r_{x_1y_1}$ = the correlation of Recall A and Recognition A.

$r_{x_2y_2}$ = the correlation of Recall B and Recognition B.

$r_{x_1x_2}$ = the reliability of the recall type.

$r_{y_1y_2}$ = the reliability of the recognition type.

The relatively higher correlations obtained when the raw correlations are corrected for attenuation indicate the presence of a high degree of identity between the criterion and the various recognition types. The r's for scores made under instructions not to guess and then corrected for chance are higher for all types except the two-response, which behaves erratically in this particular as well as others. However, all the correlations are high enough to warrant the assumption that we are measuring the same abilities with the recognition types as with the recall type of examination. This would seem to be in direct contradiction of the findings of the Ruch-Stoddard study, where the values of r corrected for attenuation ranged from .480 to .861, while those of the present study range from .792 to .988, falling below .90 in but two cases. However, this difference can probably be fully accounted for by the differences in

range of talent, in the number of cases used, and in the care with which the items and answers were selected.

TABLE XXV

CORRELATION OF RECALL AND RECOGNITION WHEN CORRECTED FOR ATTENUATION

GUESS	UNCORRECTED FOR CHANCE	CORRECTED FOR CHANCE
7-response	.967	.971
5-response	.974	.975
3-response	.916	.954
2-response	.945	.921
True-False	.943	.953
DO NOT GUESS	**UNCORRECTED FOR CHANCE**	**CORRECTED FOR CHANCE**
7-response	.980	.982
5-response	.953	.976
3-response	.925	.988
2-response	.838	.917
True-False	.827	.962

Comparisons based upon equal working-times. Since there is considerable variation in the amount of time necessary to take one hundred recall items and that required for one hundred items of the various recognition types (Table XXXI), it is essential in comparing the correlations of the several types to base such correlations upon a constant *unit of working-time* rather than upon the number of items answered. This has been done by means of the Spearman-Brown Formula, $r=\frac{nr}{1+(n-1)r}$, and Table XXVI presents the original correlations thus corrected, where n represents the number of test items which can be answered in the length of time required for one recall item.

It will be noted that the corrections thus obtained reduce the difference between the various types, so that when instructions are given not to guess and the scores are corrected for chance, the reliabilities per constant unit of working-time for all five types fall within twice the probable error

of the highest correlation. This would indicate that per unit of time one type is approximately as reliable as any other type. This conclusion does not obtain when the "guess" instructions are used, since the true-false and seven-response tests are significantly below the other three in reliability. In the case of the seven-response, this may be caused by confusion on the part of the subject or the fact that he becomes tired and hurried near the end of the second form. In the case of the true-false test it may be due to alteration of relative difficulties because of the type of answer inserted to make the statement false. The close agreement of most of the coefficients would seem to indicate little difference in the character of the tests.

TABLE XXVI

CORRELATIONS OF RECOGNITION A VS. RECOGNITION B, CORRECTED BY SPEARMAN-BROWN FORMULA

(n = number of items on recognition type per unit of 100 Recall items for constant working-time.)

TYPE OF TEST	UNCORRECTED FOR CHANCE		CORRECTED FOR CHANCE	
	Corrected by Spearman-Brown Formula	Original	Corrected by Spearman-Brown Formula	Original
7-response (g)[1]	.815	.800	.851	.839
7-response (n)[2]	.901	.886	.920	.907
5-response (g)	.884	.864	.917	.902
5-response (n)	.893	.862	.908	.882
3-response (g)	.871	.837	.883	.858
3-response (n)	.913	.886	.916	.890
2-response (g)	.805	.745	.900	.864
2-response (n)	.955	.859	.905	.843
True-False (g)	.729	.641	.842	.780
True-False (n)	.925	.884	.892	.837

Correlation of Recall A vs. Recall B	.950
Coefficient of Reliability	.970

[1](g) indicates the tests taken under instructions to guess.
[2](n) indicates the tests taken under instructions not to guess.

Means and variabilities of scores. The relative difficulty of the several test types is shown by the mean scores obtained on each test. On all examinations the mean scores were higher on the recognition type than on the recall, and the differences are all significant, since all the differences are at least three times the probable error of the difference. However, between the four types of responses the same amount of difference does not exist, with the possible exception of the two-response "do not guess," which is significantly higher than any other of the "do-not-guess" types. The true-false is evidently much harder than any of the other types of recognition tests. In Table XXVII the means reported under Recall A and Recall B are the mean scores of the particular group taking the respective recognition test. The mean of all ten groups pooled is given in the last row of the table.

TABLE XXVII—MEAN SCORES

TYPE OF TEST	RECALL A	RECOG. A (Uncorrected for Chance)	RECOG. A (Corrected for Chance)	RECALL B	RECOG. B (Uncorrected for Chance)	RECOG. B (Corrected for Chance)
7-response (g)[1].	25.9	50.0	41.5	26.2	39.6	32.6
7-response (n)[2].	27.6	44.9	40.0	27.6	37.2	33.1
5-response (g)..	25.7	54.2	43.4	26.9	45.5	35.4
5-response (n)..	28.0	48.8	42.3	28.6	42.1	36.4
3-response (g)..	25.6	62.2	43.6	26.1	55.5	36.6
3-response (n)..	27.4	54.1	41.9	27.5	48.2	36.1
2-response (g)..	26.7	71.7	43.6	27.4	67.2	37.1
2-response (n)..	33.4	65.1	45.8	32.3	60.3	40.2
True-False (g)..	27.4	65.8	32.3	26.8	61.3	26.0
True-False (n)..	27.5	51.0	30.8	27.6	47.6	26.8
Mean of the 2453 cases....	27.6	.....	.....	27.8	.....	.....

[1](g) indicates the tests taken under instructions to guess.
[2](n) indicates the tests taken under instructions not to guess.

It is interesting to note the effect of correction for chance upon the spread of the distribution as shown in Table XXVIII. The effect of the correction is uniformly to raise the standard deviation, indicating a decrease in the piling up around the mean. In the majority of cases the sigma is larger for the "guess" type than for the "do not guess."

TABLE XXVIII—SIGMAS

TYPE OF TEST	RECALL A	RECOG. A (Uncorrected for Chance)	RECOG. A (Corrected for Chance)	RECALL B	RECOG. B (Uncorrected for Chance)	RECOG. B (Corrected for Chance)
7-response (g)[1].	15.1	17.4	20.0	16.3	20.7	20.9
7-response (n)[2].	17.1	19.7	21.0	17.5	20.2	20.6
5-response (g)..	16.9	17.0	20.8	17.3	20.9	22.1
5-response (n)..	16.8	19.8	21.1	17.6	21.1	21.2
3-response (g)..	15.0	13.9	19.8	16.4	16.3	19.6
3-response (n)..	15.7	18.2	20.5	16.5	19.6	20.0
2-response (g)..	16.4	11.4	22.4	17.0	13.3	20.6
2-response (n)..	20.6	16.1	20.7	19.7	17.9	20.1
True-False (g)..	16.5	10.5	19.4	16.9	12.6	17.6
True-False (n)..	15.9	17.4	18.8	16.5	18.3	17.3

[1](g) indicates the tests taken under instructions to guess.
[2](n) indicates the tests taken under instructions not to guess.

Percentile working-times. Tables XXIX and XXX are presented with the idea that they may be somewhat of a guide in determining the number of items that can reasonably be expected of the class, or what percentage of the class can be expected to cover one hundred or fifty items within a given length of time. The percentiles of the pooled grades are based on over two thousand cases and should represent quite adequate standards. In the cases of the grades the number of cases ranges from four hundred to over five hundred, which should be sufficient to set reasonably safe standards.

The practice effect of having taken this particular type of test can easily be seen by comparing the times taken by pupils of the eleventh and twelfth grades on Forms A and B, respectively. In all percentiles of both grades the time taken on Form B is consistently lower than the time required for Form A. The top row gives the upper limit of time taken.

TABLE XXIX

PERCENTILES OF TIME IN MINUTES FOR RECALL TESTS: 100 ITEMS (BY GRADES)

%-ILES	FORM A				FORM B				FORM A	FORM B
	GRADE				GRADE				All Grades	All Grades
	7th	8th	11th	12th	7th	8th	11th	12th		
	44.5	43.0	44.0	44.0	45.0	47.5	40.5	40.5	44.5	45.5
90	30.2	33.4	34.5	36.1	30.3	34.3	32.9	34.0	34.0	33.2
80	27.4	30.6	31.1	32.7	26.4	30.7	30.1	31.2	30.7	30.2
70	25.5	28.3	29.2	30.5	24.5	28.4	28.1	29.7	28.0	27.9
60	23.6	26.5	27.3	28.4	22.5	26.7	25.5	27.2	26.6	25.7
50	22.0	25.1	25.0	26.7	21.1	24.9	24.4	25.5	24.8	24.2
40	20.5	23.6	23.8	25.5	20.0	22.9	23.0	24.2	23.2	22.5
30	19.3	21.9	22.0	23.0	18.7	21.1	21.6	22.6	21.0	21.0
20	17.5	19.7	20.1	21.3	17.5	19.7	20.0	21.5	19.6	19.0
10	15.3	17.6	18.2	18.7	15.9	17.3	18.2	19.5	17.4	17.3

TABLE XXX

PERCENTILES OF TIME IN MINUTES FOR RECOGNITION TYPES (200 ITEMS)

%-ILES	RESPONSE TYPES								TRUE-FALSE	
	7 (g)	7 (n)	5 (g)	5 (n)	3 (g)	3 (n)	2 (g)	2 (n)	(g)	(n)
	60.0	60.0	60.5	60.0	60.0	60.5	60.0	60.0	55.0	56.0
90	55.4	50.7	52.6	48.1	48.4	48.0	45.0	43.6	44.2	39.1
80	50.4	48.2	48.6	44.1	45.1	45.2	40.7	40.3	39.0	36.2
70	48.1	45.3	45.7	41.8	42.0	41.9	38.8	38.6	36.9	34.4
60	45.4	42.4	44.6	40.1	40.2	39.7	36.4	36.3	34.5	32.4
50	43.9	40.4	42.0	38.8	38.6	37.4	34.8	35.1	32.8	30.6
40	41.0	38.8	40.2	38.2	36.4	35.3	31.9	32.9	29.8	28.7
30	39.2	36.4	38.6	36.2	34.6	33.1	29.6	30.9	28.8	27.4
20	37.7	33.1	36.1	34.9	32.4	30.6	26.4	28.8	26.2	25.3
10	30.4	29.4	32.4	31.1	29.1	27.4	23.7	25.2	24.6	21.7

The mean times as shown in Table XXXI are based upon the pooled grades and show close agreement with the median (50th percentiles) of Table XXX.

TABLE XXXI

MEAN TIMES IN MINUTES

TYPE	TIME	NO. OF CASES
7-response (g) (200 items)	45.9	212
7-response (n) (200 items)	40.8	206
5-response (g) (200 items)	42.2	214
5-response (n) (200 items)	37.6	262
3-response (g) (200 items)	38.2	239
3-response (n) (200 items)	37.6	219
2-response (g) (200 items)	34.6	207
2-response (n) (200 items)	34.5	251
True-False (g) (200 items)	33.0	227
True-False (n) (200 items)	30.5	244
Recall A (100 items) 25.2 / Recall B (100 items) 24.8 } = 50.0	Cases 2200	

Comparisons with other investigations. A comparison of the results obtained in this study with those obtained by the Toops and the Ruch-Stoddard experiments (Tables XXXII and XXXIII) reveals rather close agreement in most of the correlations, when we take into consideration the wider range of the present study. The reliability of one hundred true-false items is found to be the lowest in all three studies, while the recall possesses the highest reliability for the same number of items. Ruch and Stoddard found rather wide variations in the reliabilities when the Spearman-Brown formula was used to estimate reliability for constant working-times, whereas in this, as well as in the Toops study, all of the recognition tests are equally reliable within one or two P.E.

The differences in the time required for a hundred-item test are appreciably higher in this study than in either of the

other two, doubtless due to differences in degrees of difficulty. The comparison of the number of questions per unit of recall-time for the several types of tests in the three studies presents interesting variations. Whereas in the three-response the two studies agree, this condition obtains in none of the other types. The explanation of these differences may lie in the ease or difficulty of the respective answers offered.

TABLE XXXII

COMPARISON OF SOME ITEMS OF THIS STUDY WITH THOSE OF TOOPS AND RUCH-STODDARD

STUDY	RECALL	5-RESPONSE	3-RESPONSE	2-RESPONSE	TRUE-FALSE
RELIABILITY OF 100 ITEMS					
Toops	.764	.715	....	....	.673
Ruch-Stoddard	.896	.886	.748	.849	.714
This study	.950	.882	.890	.843	.837
AVERAGE TIME IN MINUTES TO DO 100 ITEMS					
Toops	13.8	11.2	....	....	7.2
Ruch-Stoddard	18.7	16.0	13.5	11.4	10.2
This study	25.0	18.8	18.8	17.2	15.2
NUMBER OF QUESTIONS PER UNIT OF RECALL-TIME					
Toops	1.00	1.23	....	....	1.92
Ruch-Stoddard	1.00	1.17	1.39	1.64	1.83
This study	1.00	1.33	1.36	1.48	1.61
RELIABILITY OF FORM A-FORM B PER UNIT OF RECALL-TIME					
Toops	.764	.755	....	....	.798
Ruch-Stoddard	.896	.901	.806	.902	.820
This study	.950	.908	.916	.888	.892

TABLE XXXIII

NUMBER OF ITEMS PER UNIT OF 100 RECALL ITEMS

STUDY	RECALL	5-RESPONSE	TRUE-FALSE
Toops	100	123	192
Ruch-Stoddard	100	117	183
This Study	100	133	161

Table XXXIV is presented to show the mean times required for one hundred items of each of the eleven types used, and also the number of items which could be covered in a constant working-time. The constant used is, of course, the time required to cover one hundred recall questions.

TABLE XXXIV

NUMBER OF ITEMS OF EACH TYPE PER UNIT OF TIME FOR 100 RECALL ITEMS. (MEAN TIMES.)

TYPE	UNIT OF WORKING-TIME	NUMBER OF ITEMS
Recall	100 in 25 minutes	100
7-response (g)	100 in 24.9 minutes	110
7-response (n)	100 in 20.4 minutes	118
5-response (g)	100 in 21.1 minutes	120
5-response (n)	100 in 18.8 minutes	133
3-response (g)	100 in 19.1 minutes	132
3-response (n)	100 in 18.8 minutes	136
2-response (g)	100 in 17.3 minutes	141
2-response (n)	100 in 17.2 minutes	148
True-False (g)	100 in 16.5 minutes	151
True-False (n)	100 in 15.2 minutes	161

Paralleling an analysis made in the Ruch-Stoddard study to determine in what types of examination, if any, the actual number of right guesses exceeds the theoretical number, Table XXXV (page 87) shows the same tendency as that found in the above study in regard to the five-response. On the other hand, the results here show a very slight over-correction in the cases of the two- and three-response types. In both studies the amount of over-correction is large in the case of the true-false. The correction seems to be most adequate in the true-false, "guess," since the over-correction is but one-half a score-point. Of the ten types, four show the over-correction, and but one of these (the true-false, "do not guess") is badly over-corrected.

This may indicate the presence of some psychological factor in the true-false question when the subject is instructed to answer only those items of which he is reasonably sure. *However, the explanation may lie in the fact that the true-false is not in reality a two-response test, but partakes essentially of the nature of the single-response type and hence is logically not to be treated by the correction for chance.* No data have been advanced to support this explanation, but there is no question but that there is a real difference between the two choices actually offered in the two-response and the implied choices of the true-false. In the latter, the subject is confronted with the simple statement and must decide, not between two possible answers as presented, but the truth or falsity of the statement *as one unit.* In other words, the entire statement or question is the basis of decision in the case of the true-false, but in the two-response test the choice of responses is made in the light of the connection of each with the rest of the statement.

The actual as well as the relative excess of actual over theoretical right "guesses" in the four multiple-choice tests when subjects are instructed to guess, indicates that under such directions the correction formula does not penalize enough to account for the chance element.

The attention of the reader is invited to an important assumption which underlies the computations of Table XXXV. This assumption is not believed to be highly valid, but was merely accepted as a means of demonstrating one aspect of the problem of "guessing" in recognition tests. In brief, as shown by column *a*, it was assumed that the group of pupils knew *only* 55.4 items, and that 200 minus 55.4 were to be "guessed at." The point, that the pupils *might recognize among alternatives* the true responses when they could not *recall such responses spontaneously*, does not enter into the calculation of Table XXXV; but it is probably a fact, nevertheless, that pupils do have "fringes of knowledge" and "hazy ideas" which cast the die in doubtful cases.

If this is true, such responses are by no means "pure guesses," as the correction formula implies. The justification for the assumption basic to Table XXXV, is, therefore, not its truth (because it is extremely doubtful), but that such a hypothesis leads to the discrepancies of column *g*, thus throwing doubt on the correction formula itself.

In order to compare the work entailed in the scoring of the various kinds of objective examination, it was deemed advisable to determine the average number of minutes required by the average scorer to correct and total the scores of the various types. This was done by three separate scorers at different times and at different stages of practice. The recall type took the least amount of time on the average (2.23 minutes per booklet of one hundred items). This included the transferring of the complete score—"rights," "wrongs," and number "omitted," as well as the number of minutes required by the pupils to take the test. Of course, the time of scoring would vary somewhat from the central tendency, depending upon the number of items attempted by the group, which in turn varied directly with the grade in which the test was taken. The recognition types ranged in descending order as follows: true-false, (3.95 minutes); seven-response, (3.75); five-response, (3.62); three-response, (3.12); and two-response, (2.87). It might be mentioned in passing that all the scorers found the true-false the most difficult type to score and the recall the easiest.

IV. CONCLUSIONS

The main conclusions that have been reached as a result of this experiment may be stated briefly as follows:

(1) The evidence collected in this study points to the superiority of the method of correcting for chance $\left(S=R-\frac{W}{n-1}\right)$ over the method of counting the score as the number of right answers. Whereas the evidence is

TABLE XXXV

Analysis of Validity of Corrections for Chance

Type of Examination	(*a*) Means of Uncorrected Scores A+B	(*b*) Means of Scores Corrected for Chance	(*c*) Differences (*a* −*b*)	(*d*) No. of Items to be "Guessed" (200—Recall Means)	(*e*) Theoretical No. of Right "Guesses" 144.6+n	(*f*) Actual No. of Right "Guesses" (*a*) −55.4	(*g*) Excess of Actual Right "Guesses" over Theoretical
Recall	55.4	55.4[1]	00.0	144.6	....[1]	00.0	.....[1]
7-response (g)	89.6	74.1	15.5	144.6	20.7	34.2	+13.5
7-response (n)	82.1	73.1	9.0	144.6	20.7	26.7	+ 6.0
5-response (g)	99.7	78.8	20.9	144.6	28.9	44.3	+14.4
5-response (n)	90.9	78.7	12.2	144.6	28.9	35.5	+ 6.6
3-response (g)	118.8	80.1	38.7	144.6	48.2	63.4	+15.2
3-response (n)	102.3	78.0	24.3	144.6	48.2	46.9	− 1.3
2-response (g)	138.9	80.7	58.2	144.6	72.3	83.5	+11.2
2-response (n)	125.4	85.9	39.5	144.6	72.3	70.0	− 2.3
True-false (g)	127.2	59.3	67.9	144.6	72.3	71.8	− 0.5
True-false (n)	108.6	57.6	51.0	144.6	72.3	53.2	−19.1

[1]Chance element does not enter.

not overwhelmingly on the side of correction, it is nevertheless sufficiently strong to warrant the continued use of such correction until further study shall develop a more accurate technique.

(2) Of the six types of objective examinations used in this study, the recall is the most reliable. Of the recognition types, the five-response and seven-response tests are about equal in ranking as to reliability, while the two-response and true-false types are almost uniformly the most unreliable.

(3) The evidence is quite clearly indicative that it is better to instruct the subject to answer only those questions of which he is reasonably sure rather than to attempt all items presented.

(4) A comparison of the mean scores made arranges the six types in the following order of decreasing difficulty: recall, true-false, seven-response, three-response, five-response, and two-response.

(5) The mean times required by the pupil to take the various types of examinations rank the tests in the following order as regards increasing time consumption: true-false, two-response, three-response, five-response, seven-response, and recall.

(6) For scoring, the true-false takes the greatest amount of time, the recall requires the least time, and the others are ranged in descending order, as follows: seven-, five-, three-, and two-response types. This order would obtain only when the method of recording answers (by numbers in lieu of underscoring) was used.

CHAPTER V

A STUDY OF THE TECHNIQUE OF MATCHING TESTS[1]

I. MATERIALS AND METHOD

Purpose of the investigation. The primary purpose of the present study was to determine experimentally the optimum number of test items per group in matching tests. By optimum is meant that grouping which would meet the following conditions:

(*a*) Reduce the amount of chance of guessing to a negligible degree, if possible, but

(*b*) not increase the amount of time required to match the test elements to an uneconomical amount, or

(*c*) make the test unwieldy and unattractive to the pupils.

Any thought of checking up the knowledge of pupils about dates and characters in history was definitely a secondary consideration, especially since no claims to high validity or social usefulness can be made for much of the content of these matching tests.

The tests employed. Twelve tests were used, as follows:

SERIES I

TEST 1—Matching Test:	Dates and Events.	Grouping by 5's (I)
TEST 2—Matching Test:	Dates and Events.	Grouping by 5's (II)
TEST 3—Matching Test:	Dates and Events.	Grouping by 10's
TEST 4—Matching Test:	Dates and Events.	Grouping by 15's
TEST 5—Matching Test:	Dates and Events.	Grouping by 20's
TEST 6—Matching Test:	Dates and Events.	Grouping by 30's

[1]The matching tests used in this study were developed by Miss Nell Maupin and Mr. John R. Murdock. The latter is chiefly responsible for the scoring and statistical work.

SERIES II

TEST 7—Matching Test: Men and Characterizations. Grouping by 5's
TEST 8—Matching Test: Men and Characterizations. Grouping by 10's
TEST 9—Matching Test: Men and Characterizations. Grouping by 15's
TEST 10—Matching Test: Men and Characterizations. Grouping by 20's
TEST 11—Matching Test: Men and Characterizations. Grouping by 30's

SERIES III

TEST 12—Matching Test: Men and Characterizations. (Completion Form)

In both Series I and Series II, 120 items were selected. These were then broken by chance into two forms, designated as Form A and Form B. The next step was that of subdividing the 60 items of each form into "Sections" or groupings of 5, 10, 15, 20, and 30 items each. No matter which grouping arrangement was employed, both Form A and Form B were printed as a single 4-page booklet, Form A invariably appearing first.

The form and content of these matching tests will be made clearer by an examination of the specimen pages reprinted here.

The first page of Test 1 of Series I is reproduced in full. The actual items of the first page of Test 2 of Series I are also shown.

The first page of Test 7, Series II is reproduced in full except for the customary blanks calling for information regarding the students. The actual items of the first page of Tests 8 and 11 of Series II are also shown.

The first page of Test 12, Series III, is reproduced in full except for the student's information blanks.

SERIES I TEST 1

MATCHING TEST: DATES AND EVENTS

For experimental purposes only
Commonwealth Investigation of Examinations in the Social Sciences
University of Iowa

Name..... Age..... Grade.....

Sex..... School..... .....
(City) (State)

Date of this Test . ..... Date of Birth.....
(Month) (Day) Year)

Directions: Match the *dates* and *events* below by writing the correct date in the parenthesis at the left of each event. Notice that the first item is already filled in correctly. Work as fast as you can without making mistakes.

FORM A

DATES		EVENTS
		Section 1
1453	(*1534*)	Discovery of the St. Lawrence River by Cartier
1492	( )	Sea trade with India established by Vasco da Gama
1498	( )	Defeat of the Spanish Armada by the English
1534	( )	Capture of Constantinople by the Turks
1588	( )	First voyage of Columbus across the Atlantic Ocean
		Section 2
1608	( )	Introduction of slavery into Virginia
1619	( )	Founding of Pennsylvania by William Penn
1643	( )	Settlement of Quebec by the French
1682	( )	Meeting of the Albany Congress
1754	( )	Meeting of the New England Confederation
		Section 3
1765	( )	Inauguration of Washington as the first president
1776	( )	Controversy over the Stamp Act
1781	( )	Passage of the Northwest Ordinance
1787	( )	The Declaration of Independence
1789	( )	Ratification of the Articles of Confederation
		Section 4
1790	( )	Decision of the Supreme Court in the case of Marbury vs. Madison
1793	( )	Passage of the Embargo Act under Jefferson
1800	( )	Birth year of the factory system and the first tariff law
1803	( )	Establishment of Washington, D.C., as capital of United States
1807	( )	Invention of the cotton gin by Eli Whitney

Section 5

1814 ( ) Opening of the Erie Canal
1819 ( ) Beginning of the era of railroad building in the U. S.
1820 ( ) Meeting of the Hartford Convention
1825 ( ) Purchase of Florida from Spain
1830 ( ) Passage of the Missouri Compromise

Section 6

1832 ( ) Annexation of Texas by the United States
1845 ( ) Passage of the Compromise on slavery championed by Clay
1846 ( ) Discovery of gold in California
1848 ( ) Controversy over nullification heads up in South Carolina
1850 ( ) Invention of the sewing machine by Howe

Go Right on to Page 2

SERIES I, TEST 2

FORM A

DATES		EVENTS

Section 1

1453 (*1872*) Settlement of the Alabama claims at Geneva
1781 ( ) Beginning of the era of railroad building in the U. S.
1830 ( ) Capture of Constantinople by the Turks
1872 ( ) Passage of the Pure Food and Drug Act
1906 ( ) Ratification of the Articles of Confederation

Section 2

1492 ( ) Controversy over Nullification heads up in South Carolina
1787 ( ) Invention of the arc light for streets and parks
1832 ( ) Assembly of the Second International Peace Congress at The Hague
1878 ( ) First voyage of Columbus across the Atlantic Ocean
1907 ( ) Passage of the Northwest Ordinance

Section 3

1498 ( ) Inauguration of Washington as the first president
1789 ( ) Touring the world by the United States fleet
1845 ( ) Sea trade with India established by Vasco da Gama
1883 ( ) Annexation of Texas by the United States
1908 ( ) Passage of the Civil Service Reform Act

Section 4

1534 ( ) Organization of the American Federation of Labor
1790 ( ) Law passed requiring candidates for Congress to publish campaign expenses
1846 ( ) Discovery of the St. Lawrence River by Cartier
1881 ( ) Birth year of the factory system and the first tariff law
1910 ( ) Invention of the sewing machine by Howe

Section 5

1588 ( ) Discovery of gold in California
1793 ( ) Passage of the Interstate Commerce Act
1848 ( ) Passage of the Federal Reserve Banking Act
1887 ( ) Defeat of the Spanish Armada by the English
1913 ( ) Invention of the cotton gin by Eli Whitney

Section 6

1608 ( ) Establishment of Washington, D.C., as capital of the United States
1800 ( ) The Pullman strike—injunction issued
1850 ( ) Settlement of Quebec by the French
1894 ( ) Beginning of the World War
1914 ( ) Passage of the Compromise on slavery championed by Clay

Go Right on to Page 2

SERIES II TEST 7

MATCHING TEST: MEN AND CHARACTERIZATIONS

Directions: Read each *characterizing phrase* and then find the *man* at the left whom the phrase fits best. Record the *number* of the proper man in the parenthesis in front of each phrase. Notice that the first item is already filled in correctly. ***Each phrase must be matched with a man in the same section.*** Work as fast as you can without making mistakes.

FORM A

MEN		CHARACTERIZING PHRASE
		Section 1
1. Thomas H. Benton	(*5*)	Author of the Declaration of Independence
2. Thaddeus Stevens	( )	For thirty years a senator from Missouri
3. George B. McClellan	( )	An immigrant who worked for political reform
4. Carl Schurz	( )	Leader of Union Army in Peninsula Campaign
5. Thomas Jefferson	( )	Congressman demanding harsh treatment of South
		Section 2
6. Miles Standish	( )	Discoverer of the New World for Spain
7. De Witt Clinton	( )	Spent a fortune to found a colony in America
8. Charles Sumner	( )	Military man of Plymouth, told of by Longfellow
9. Sir Walter Raleigh	( )	Massachusetts senator denouncing "Crime Against Kansas"
10. Christopher Columbus	( )	Governor of New York—promoted the Erie Canal
		Section 3
11. Vasco de Balboa	( )	Governor of Plymouth Colony and Pilgrim leader
12. David Wilmot	( )	Discovered the South Sea, or Pacific Ocean
13. Woodrow Wilson	( )	Offered a "Proviso" concerning slave territory
14. William Bradford	( )	Portuguese explorer who rounded Africa to India
15. Vasco de Gama	( )	The "Great War President"—League of Nations
		Section 4
16. William T. Sherman	( )	U. S. agent to France during X-Y-Z affair
17. Cyrus W. Fields	( )	Invented cylindrical newspaper printing press
18. John Winthrop	( )	Northern general who marched through Georgia
19. Richard Hoe	( )	Laid the first successful Atlantic cable
20. Charles C. Pinckney	( )	Puritan Governor of Massachusetts Bay Colony
		Section 5
21. James J. Hill	( )	Fifth President—"Era of Good Feeling"
22. William McKinley	( )	Orator who denounced "Writs of Assistance"
23. James Monroe	( )	Railroad "King"—builder of Northwest

24. Henry Clay	( )	Third "Martyr President"—shot by anarchist
25. James Otis	( )	Kentucky statesman famous for compromises

Section 6

26. John Jacob Astor	( )	A South Carolina advocate of nullification
27. Andrew Jackson	( )	The "Little Giant" debating with Lincoln
28. John C. Calhoun	( )	Founded a fur-trading company in Oregon
29. George H. Meade	( )	Hero of New Orleans, first president from West
30. Stephen A. Douglas	( )	Northern general who won at Gettysburg

Go Right on to Page 2

SERIES II, TEST 8

FORM A

MEN		CHARACTERIZING PHRASE

Section 1

1. Thomas H. Benton	(*5*)	Author of the Declaration of Independence
2. Thaddeus Stevens	( )	For thirty years a senator from Missouri
3. George B. McClellan	( )	An immigrant who worked for political reform
4. Carl Schurz	( )	Leader of Union Army in Peninsula Campaign
5. Thomas Jefferson	( )	Congressman demanding harsh treatment of South
6. Miles Standish	( )	Discoverer of the New World for Spain
7. De Witt Clinton	( )	Spent a fortune to found a colony in America
8. Charles Sumner	( )	Military man of Plymouth, told of by Longfellow
9. Sir Walter Raleigh	( )	Massachusetts senator denouncing "Crime Against Kansas"
10. Christopher Columbus	( )	Governor of New York—promoted the Erie Canal

Section 2

11. Vasco de Balboa	( )	Governor of Plymouth Colony and Pilgrim leader
12. David Wilmot	( )	Discovered the South Sea, or Pacific Ocean
13. Woodrow Wilson	( )	Offered a "Proviso" concerning slave territory
14. William Bradford	( )	Portuguese explorer who rounded Africa to India
15. Vasco da Gama	( )	The "Great War President"—League of Nations
16. William T. Sherman	( )	U. S. agent to France during X-Y-Z affair
17. Cyrus W. Fields	( )	Invented cylindrical newspaper printing press
18. John Winthrop	( )	Northern general who marched through Georgia
19. Richard Hoe	( )	Laid the first successful Atlantic cable
20. Charles C. Pinckney	( )	Puritan Governor of Massachusetts Bay Colony

Section 3

21. James J. Hill	( )	Fifth President—"Era of Good Feeling"
22. William McKinley	( )	Orator who denounced "Writs of Assistance"
23. James Monroe	( )	Railroad "King"—builder of Northwest
24. Henry Clay	( )	Third "Martyr President"—shot by anarchist
25. James Otis	( )	Kentucky statesman famous for compromises
26. John Jacob Astor	( )	A South Carolina advocate of nullification
27. Andrew Jackson	( )	The "Little Giant" debating with Lincoln
28. John C. Calhoun	( )	Founded a fur trading company in Oregon
29. George H. Meade	( )	Hero of New Orleans, first president from West
30. Stephen A. Douglas	( )	Northern general who won at Gettysburg

Go Right on to Page 2

SERIES II, TEST 11

FORM A

MEN		CHARACTERIZING PHRASE
	Section 1	
1. Thomas H. Benton	(5)	Author of the Declaration of Independence
2. Thaddeus Stevens	( )	An immigrant who worked for political reform
3. George B. McClellan	( )	Invented cylindrical newspaper printing press
4. Carl Schurz	( )	Fifth Pesident—"Era of Good Feeling"
5. Thomas Jefferson	( )	Discoverer of the New World for Spain
6. Miles Standish	( )	Spent a fortune to found a colony in America
7. De Witt Clinton	( )	Founded a fur trading company in Oregon
8. Charles Sumner	( )	Military man of Plymouth, told of by Longfellow
9. Sir Walter Raleigh	( )	Massachusetts senator denouncing "Crime Against Kansas"
10. Christopher Columbus	( )	Congressman demanding harsh treatment of South
11. Vasco de Balboa	( )	Hero of New Orleans, first president from West
12. David Wilmot	( )	Puritan Governor of Massachusetts Bay Colony
13. Woodrow Wilson	( )	For thirty years a senator from Missouri
14. William Bradford	( )	Governor of Plymouth Colony and Pilgrim leader
15. Vasco da Gama	( )	Portuguese explorer who rounded Africa to India
16. William T. Sherman	( )	Orator who denounced "Writs of Assistance"
17. Cyrus W. Fields	( )	U. S. agent to France during X-Y-Z affair
18. John Winthrop	( )	Northern general who marched through Georgia
19. Richard Hoe	( )	Laid the first successful Atlantic cable
20. Charles C. Pinckney	( )	Leader of Union Army in Peninsula Campaign
21. James J. Hill	( )	Railroad "King"—builder of Northwest
22. William McKinley	( )	Third "Martyr President"—shot by anarchist
23. James Monroe	( )	The "Little Giant" debating with Lincoln
24. Henry Clay	( )	Governor of New York—promoted the Erie Canal
25. James Otis	( )	Northern general who won at Gettysburg
26. John Jacob Astor	( )	Discovered the South Sea, or Pacific Ocean
27. Andrew Jackson	( )	Offered a "Proviso" concerning slave territory
28. John C. Calhoun	( )	A South Carolina advocate of nullification
29. George H. Meade	( )	The "Great War President"—League of Nations
30. Stephen A. Douglas	( )	Kentucky statesman famous for compromises

Go Right on to Page 2

SERIES III TEST 12

MATCHING TEST: MEN AND CHARACTERIZATIONS

Directions: Write on each dotted line at the left the name of the *man* whom the *characterizing phrase* at the right fits best. Notice that the first item is already filled in correctly. Work as fast as you can without making mistakes.

FORM A

MEN	CHARACTERIZING PHRASE
1. *Thomas Jefferson*	Author of the Declaration of Independence
2.	An immigrant who worked for political reform
3.	Invented cylindrical newspaper printing press
4.	Fifth President—"Era of Good Feeling"

5. ..... Discoverer of the New World for Spain
6. ..... Spent a fortune to found a colony in America
7. ..... Founded a fur trading company in Oregon
8. ..... Military man of Plymouth, told of by Longfellow
9. ..... Massachusetts senator denouncing "Crime Against Kansas"
10. ..... Congressman demanding harsh treatment of South
11. ..... Hero of New Orleans, first president from West
12. ..... Puritan Governor of Massachusetts Bay Colony
13. ..... For thirty years a senator from Missouri
14. ..... Governor of Plymouth Colony and Pilgrim leader
15. ..... Portuguese explorer who rounded Africa to India
16. ..... Orator who denounced "Writs of Assistance"
17. ..... U. S. agent to France during X-Y-Z affair
18. ..... Northern general who marched through Georgia
19. ..... Laid the first successful Atlantic cable
20. ..... Leader of Union Army in Peninsula Campaign
21. ..... Railroad "King"—builder of Northwest
22. ..... Third "Martyr President"—shot by anarchist
23. ..... The "Little Giant" debating with Lincoln
24. ..... Governor of New York—promoted the Erie Canal

Go Right on to Page 2

A few comments are needed in connection with the difference between Tests 1 and 2 of Series I, since both are groupings by fives. Two 5-groupings were needed as a check on an unavoidable situation arising through the successive poolings of "sections" of dates; viz., the fact that in the course of such repeated poolings to produce the 10-, 15-, 20-, and 30-group arrangements, dates come to lie closer together, *on the average*, in the larger groupings. This makes the greater difficulty of the larger groupings depend not only upon (*a*) reduced operation of chance and guessing, but also (*b*) upon a constantly increasing *fineness of discrimination*. Since our teaching of most dates is approximate (or it should be approximate), confusions of near-lying dates in the coarse groupings are bound to arise. For this reason two different arrangements of the 5-grouping form of the tests on dates were made, as follows:

Series I, Test 1, involving the maximum amount of discrimination possible, since the dates within a group are separated by the smallest average number of years possible.

Series I, Test 2, involving the minimum amount of dis-

crimination possible, since the dates within a group are separated by the largest average number of years possible.

To be specific, Section 1 of Test 1, Series I, shows a range of dates from 1453 to 1588, while Section 1 of Test 2, Series I, shows a range of dates from 1453 to 1906.

Series II does not involve such a systematic increase in difficulty due to pooling, other than increased difficulty on account of reduction of chance effects.

Test 12 of Series III is not a matching test at all, but is to be thought of as a completion form of the materials in Series II. Test 12 was used primarily as a control test in order to check up on the influence of guessing in the recognition (matching) forms of the same facts.

II. RESULTS OF THE INVESTIGATION

The statistical findings. Tables XXXVI to XLI present the summaries of the data secured in the three series of matching tests.

TABLE XXXVI

RESULTS FOR EIGHTH-GRADE PUPILS ON SERIES I, TESTS 1 TO 6

	TEST NO. 1	TEST NO. 2	TEST NO. 3	TEST NO. 4	TEST NO. 5	TEST NO. 6
MEANS						
Form A....	9.70	15.13	6.39	6.11	5.20	5.00
Form B....	7.80	14.26	4.40	4.20	3.20	3.12
STANDARD DEVIATIONS						
Form A....	7.15	14.93	5.04	4.87	3.98	3.21
Form B....	6.72	13.74	4.92	3.80	3.83	3.00
r_{AB}	.86	.96	.74	.84	.79	.76
PUPILS	129	128	124	121	127	124

TABLE XXXVII

RESULTS FOR HIGH-SCHOOL SENIORS ON SERIES I, TESTS 1 TO 6

	TEST NO. 1	TEST NO. 2	TEST NO. 3	TEST NO. 4	TEST NO. 5	TEST NO. 6
MEANS						
Form A....	16.80	32.10	11.46	10.40	9.50	9.90
Form B....	13.66	29.00	8.70	8.20	6.80	6.90
STANDARD DEVIATIONS						
Form A....	8.42	15.87	6.44	5.46	6.67	6.32
Form B....	8.59	17.48	7.05	5.65	5.76	5.54
r_{AB}	.90	.93	.86	.80	.88	.91
PUPILS	130	133	121	124	125	124

TABLE XXXVIII

RESULTS FOR SERIES I, TESTS 1 TO 6, FOR ALL PUPILS

	TEST NO. 1	TEST NO. 2	TEST NO. 3	TEST NO. 4	TEST NO. 5	TEST NO. 6
MEANS						
Form A....	13.28	23.81	8.90	8.29	7.39	7.46
Form B....	10.76	21.79	6.55	6.27	5.60	5.34
STANDARD DEVIATIONS						
Form A....	8.53	17.03	5.87	5.29	5.35	5.20
Form B....	8.36	17.47	5.98	5.07	5.16	4.45
r_{AB}	.90	.95	.84	.82	.85	.85
PUPILS	259	261	245	245	252	248

TABLE XXXIX

RESULTS FOR EIGHTH-GRADE PUPILS ON SERIES II, TESTS 7 TO 12

	TEST NO. 7	TEST NO. 8	TEST NO. 9	TEST NO. 10	TEST NO. 11	TEST NO. 12
MEANS						
Form A....	35.0	26.0	22.1	20.0	16.3	13.8
Form B....	39.0	28.5	24.7	20.6	16.8	13.0
STANDARD DEVIATIONS						
Form A....	13.12	10.58	10.06	10.19	8.53	8.05
Form B....	14.52	13.21	12.40	11.95	9.55	8.47
r_{AB}	.93	.77	.89	.90	.88	.92
PUPILS	161	164	168	159	160	127

TABLE XL

RESULTS FOR HIGH-SCHOOL SENIORS ON SERIES II, TESTS 7 TO 12

	TEST NO. 7	TEST NO. 8	TEST NO. 9	TEST NO. 10	TEST NO. 11	TEST NO. 12
MEANS						
Form A....	44.7	36.7	32.3	29.9	26.2	17.0
Form B....	48.0	40.2	34.1	28.2	26.7	16.0
STANDARD DEVIATIONS						
Form A....	11.23	12.73	12.90	12.56	10.43	9.19
Form B....	12.56	14.48	14.88	14.66	11.95	9.84
r_{AB}	.94	.93	.92	.98	.95	.90
PUPILS	148	151	146	145	146	187

TABLE XLI

RESULTS FOR SERIES II, TESTS 7 TO 12, FOR ALL PUPILS

	TEST NO. 7	TEST NO. 8	TEST NO. 9	TEST NO. 10	TEST NO. 11	TEST NO. 12
MEANS						
Form A....	39.77	31.12	26.86	24.35	21.02	15.53
Form B....	43.57	34.11	29.09	24.14	21.62	15.05
STANDARD DEVIATIONS						
Form A....	13.20	13.06	12.53	12.28	10.70	8.87
Form B....	14.37	15.24	14.36	13.85	11.76	9.43
r_{AB}						
	.94	.94	.91	.92	.91	.91
PUPILS						
	309	315	314	304	306	314

III. DISCUSSION OF THE RESULTS

Matching of dates and events (Series I). Tables XXXVI, XXXVII, and XXXVIII present the data secured by means of the date-event tests. Tables XXXVI and XXXVII present the results for grades eight and twelve separately, and Table XXXVIII gives the combined facts. Since the larger populations of the pooled results of Table XXXVIII make for statistical stability of the quantitative results, and since, further, the results for the two grades separately are in quite general accord with the combined figures of Table XXXVIII, the discussion may well be confined to this one table.

With reference to mean scores (averages), it will be noted that the 10-grouping is distinctly more difficult than the 5-grouping. (Cf. Tests 1 and 3.) However, the differences between the 10-, 15-, 20-, and 30-groupings (Tests 3, 4, 5,

and 6), although probably statistically significant in some cases, *would seem to have no practical importance.* This means that groupings coarser than 10 or 15 offer very little improvement with respect to the minimization of guessing over grouping by 10's. The slight gain which is made possible through coarse groups is probably more than offset by the extra time that is needed for the administration of long units.[1]

These results indicate rather definitely that grouping by 10's is to be preferred over grouping by 5's as a device for eliminating guessing or chance effects.

Test 2 of Series I (minimum discrimination) is by far the easiest, as was to be expected. Since, however, Test 1 is probably more comparable with Tests 3 to 6, it is safer to trust Test 1 for grouping comparisons.

The variabilities of the scores (standard deviations) show about the same facts as were found in the case of mean scores.

The reliability coefficients are more difficult of interpretation. On the whole, the 5-groupings seem to be definitely superior, the obtained differences being statistically significant between the 5-groupings and any of the larger groupings. The reason for this may be found to lie in the fact that the 10-, 15-, etc., groupings are *in effect* much shorter tests (as shown by the means); i.e., the pupils responded to fewer items because of the successively increasing difficulty. This effect in test scores is well known and may have, in this case, more than overbalanced the gains in reliability that are to be expected, a priori, with reduced chance effects in coarse groupings.

At any rate, this hypothesis is in harmony with the fact that Test 2 showed a higher reliability than Test 1, the

[1]No data appear here on the relative amounts of time needed for completing the entire 120 items for each of the six grouping arrangements. As might be expected, the 30-grouping required markedly more time than the 5- or 10-grouping, since the pupils had to spend more time in searching out the paired responses. It is planned to publish exact data on the working times of these tests at a later date.

magnitudes of the mean scores showing differences in the same direction. Thus:

	Means	r_{AB}
TEST 1 (Maximum Discrimination)		
Form A	13.23	
Form B	10.76	.90
TEST 2 (Minimum Discrimination)		
Form A	23.31	
Form B	21.79	.95
TEST 3		
Form A	8.90	
Form B	6.55	.84
Etc.		

It is unfortunate that these matching tests proved to be so difficult as to shorten the effective lengths of the test to a relatively small percentage of the printed lengths of the tests. Had the average performance been nearer the theoretically ideal score (about 30, i.e., 50 per cent of the items in a form), the results might have given a clearer picture on this matter of relative reliabilities of different groupings.[1]

[1]The temptation is presented to attempt to harmonize these differences in the reliability coefficients by means of the Spearman-Brown prophecy formula, $r_{nn}=\frac{nr}{1+(n-1)r}$, using the standard deviations of the several tests of the series as measures of the effective lengths of the tests. As a matter of fact, could the assumptions of the Spearman-Brown formula be proved to be met by our data, the difficulties in our explanations would largely disappear.

If we take Test 2 as a point of reference, and by using the r's of the other tests as bases for prediction in turn, setting n equal in turn to the ratios of the average standard deviation of Test 2 to the average standard deviation of each other test, we obtain the following values of r_{nn} for Tests 1, 3, 4, 5, and 6:

TEST	AVERAGE S.D.	RATIO OF S.D. OF TEST 2 TO S.D. OF OTHER TESTS	r_{AB}	r_{nn} (Estimated)
1	8.44	17.25/8.44 = 2.04	.90	.95
3	5.92	17.25/5.92 = 2.91	.84	.94
4	5.18	17.25/5.18 = 3.33	.82	.94
5	5.25	17.25/5.25 = 3.29	.85	.95
6	4.82	17.25/4.82 = 3.58	.85	.95

The calculations in the right hand column must be taken for what they are worth in the light of the assumptions made. The closeness of all the values to .95 (the basal r in our point of reference, i. e., the reliability of Test 2) is striking and may provide some support for our hypothesis that the tests vary so much in *effective* length as to overshadow unreliability effects attributable solely to chance or guessing.

Matching of men and characterizations. Tables XXXIX, XL, and XLI present results for the second series of matching tests (Series II, Tests 7 to 11) and for the completion form of the same materials (Series III, Test 12). Table XLI will serve as a working summary of all three tables.

All of the interpretations previously advanced for the date-event matching exercises seem to hold fairly well for these men-characterization groupings. Fortunately, the results from Table XLI are even less obscured by interfering factors than was the case with Table XXXVIII. In general, the mean scores show that more or less guessing takes place all the way through the series, the behavior of Tests 9 to 11 showing evidence that guessing probably is not completely eliminated by any of the arrangements that were tried out.

If we compare Tests 11 and 12, we find that the recall (completion) form is considerably more difficult than the 30-grouping recognition form. The difference may not be due to chance effects, i.e., guessing, but may show merely that *it is easier to recognize facts than it is to recall such spontaneously.*

The variabilities do not change greatly from grouping to grouping, and the *r's* are generally higher than for the date-event series, presumably because the tests are longer *in effect.* On the whole, the behavior of the reliability coefficients supports the hypothesis advanced in connection with the correlations of the preceding series.

No distinct "point of diminishing returns" can be located, here, from inspection of the means, which will serve to indicate exactly how coarse a grouping should be used. Since the evidence from the variabilities, reliabilities, and relative times would seem to be more significant than that from the means alone, it is probably safe to hold to the previous recommendation that groupings of 10 to 15, or at most 20, items are a fair compromise between the theoretical situation and the administrative situation.

IV. SUMMARY AND CONCLUSIONS

(1) The data presented in this chapter, although not entirely unambiguous, seem to indicate that groupings of 10 to 15 items in matching tests are as safe a practice as any which can be suggested at present.

(2) Guessing effects are not completely eliminated even in groupings of 30 items.

(3) Groupings of 15 to 30 are relatively less reliable than groupings of 5 and 10, but these differences may be secondary effects introduced by the fact that the increased difficulty of the tests, arising from coarse groupings, also has the effect of lessening the *effective length* of the tests, with resulting lessening of reliability.

CHAPTER VI

CRITICAL STUDIES OF THE STANDARDIZED TESTS IN THE SOCIAL STUDIES FOR HIGH SCHOOLS[1]

I. PURPOSES, TEST MATERIALS, AND METHOD

Purposes of the investigation. This study was planned to determine the intercorrelations and self-correlations of a number of tests in the social studies with a view to answering such questions, as:

(*a*) To what degree do the several available tests measure the same abilities?
(*b*) How reliable are the various tests?
(*c*) Which of the available tests are the most satisfactory?

History tests have been notoriously unsatisfactory, as is shown by the fact that no tests in this subject have ever attained a real foothold with school people. Testing in the social studies has been sporadic and unsatisfactory—teachers giving as the reason the lack of worth-while tests. The writer has secured the opinion of many dozens of superintendents and teachers, and these were practically of one mind in saying that the development of standardized tests in history has lagged far behind the advances in other fields. One of the largest test-publishing houses, when interviewed, stated that they believed that no history test will find widespread favor at present.

The reasons for the general backwardness of measurement in the field of history are not hard to find. The test situa-

[1]An abstract of a thesis submitted to the Graduate College of the State University of Iowa in partial fulfillment of the requirements for the M.A. degree, by John R. Murdock, June, 1925.

tion merely reflects the more general condition of lack of definite programs of instruction in the social subjects.

The tests and scales studied. Since all tests and all forms of each test were to be given to the same pupils, it was necessary to limit somewhat the number of tests and test sittings. Ten sittings were thought to be about the desired maximum. A few of the less well-known tests had, therefore, to be omitted. The aim was to select all of the most recent tests and all of the earlier tests which had earned any marked degree of recognition.

The ten tests finally included were as follows:

1. Barr, *Diagnostic Tests in American History, Series 2A*
2. Barr, *Diagnostic Tests in American History, Series 2B*
3. Kepner, *Background Test in Social Science, Form A*
4. Kepner, *Background Test in Social Science, Form B*
5. Gregory, *Tests in American History, Test III, Form A*
6. Gregory, *Tests in American History, Test III, Form B*
7. Pressey and Richards, *Test for the Understanding of American History* (one form only)
8. Van Wagenen, *Reading Scales, History Scale A*
9. Van Wagenen, *Reading Scales, History Scale B*
10. Van Wagenen, *American History Scales, Information Scale S3*

Where but a single form of a test was available (Nos. 7 and 10), the calculation of reliability coefficients was carried out by breaking the tests into chance halves. The tests were given in ten sittings of one class period each. Usually one or two days intervened between sittings. The actual testing was done by Mr. J. B. McGregor, later one of the investigational staff of the present studies.

The pupils tested. The main experimental group was composed of 240 juniors and seniors from a typical high school, that of Mason City, Iowa. The reliability coefficients were figured, however, on the basis of somewhat larger numbers of pupils, since several other smaller schools were also tested.

II. THE STATISTICAL TREATMENT OF RESULTS

Reliability coefficients. Table XLII gives the reliability coefficients, together with the means, sigmas, and average working-times for each of the ten tests. The starred values represent coefficients obtained from correlation of odd- vs. even-numbered items "stepped up" by means of the Spearman-Brown prophecy formula.

The means are usually slightly higher for Form B than for Form A, the differences being about what one might expect as practice effects arising from giving the tests in the order of Form A followed by Form B. Standard deviations (sigmas) are in all cases equal to nearest integral value.

Evidence from means and sigmas supports the conclusion that all the tests show practical equivalence of forms.

TABLE XLII

RELIABILITY COEFFICIENTS FOR TEN TEST FORMS

	GREGORY	BARR	P. & R.	V.W.S.–3	KEPNER	V.W. HIST. R.
N	290	279	296	225	215	217
r	.79	.71	.89*	.76*	.79	.57
			MEANS			
A	38.7	47.5	56.2	75.5	48.3	82.1
B	41.8	48.8			50.5	82.3
			STANDARD DEVIATIONS			
A	16	12	14	8	8	7
B	16	12			8	7
			AVERAGE TIME PER TEST IN MINUTES			
A	35	51	23	18	28	33
B	34	46			24	27

The outstanding fact of Table XLIII is the very low average correlation of each test or form against a composite of the remaining nine. The Gregory, Pressey-Richards, and Kepner tests are approximately equally good from this point of view. The Barr and Van Wagenen S-3 make a somewhat

poorer showing. The Van Wagenen History Reading Test shows the lowest average value, as might be expected since it is not intended as a means of mastery of historical content, but is, as its name implies, a measure of ability to read historical materials.

The Kepner Social Background Test presents a situation to be explained, since it also is not a measure of mastery of courses in history per se. This test, however, seems to be about as good a test of factual knowledge as any of the others which actually purport to do this thing. The results raise the question whether the Kepner tests do measure abilities other than mastery of content of history courses. Table XLIV throws additional light on this question.

Coefficients corrected for attenuation. Table XLIV shows the intercorrelations of the six different history tests after correction for attenuation due to errors of measurement, by the formula:

$$r_{\infty x \infty y} = \frac{\sqrt[4]{r_{x_1y_1}\ r_{x_1y_2}\ r_{x_2y_1}\ r_{x_2y_2}}}{\sqrt{r_{x_1x_2}\ r_{y_1y_2}}}$$

In the case of the two tests (Pressey-Richards and Van Wagenen S-3) existing in but one form, the formula was simplified. The notations have meaning as follows:

$r_{\infty x \infty y}$ = estimated "true" correlation, i.e., the correlation which theoretically would be expected if the actual raw correlations were not "diluted" by errors of measurement.

x_1 = Form A of any test
x_2 = Form B of any test
y_1 = Form A of any other test
y_2 = Form B of any other test

The numerator terms are the correlations for Table XLIII and the denominators are obtained from Table XLII.

TABLE XLIII

Intercorrelations of the Ten Test Forms

Test Form	1	2	3	4	5	6	7	8	9	10	Average
1. Gregory A	.....	.79	.59	.69	.72	.68	.76	.71	.49	.51	.66
2. Gregory B	.79	.....	.52	.61	.72	.70	.72	.69	.49	.45	.63
3. Barr A	.59	.52	.....	.71	.61	.52	.60	.53	.35	.44	.54
4. Barr B	.69	.61	.71	.....	.69	.57	.65	.56	.53	.52	.61
5. P. and R	.72	.72	.61	.69	.....	.67	.82	.71	.56	.55	.67
6. V. W. S-3	.68	.70	.52	.57	.67	.....	.72	.66	.45	.44	.60
7. Kepner A	.76	.72	.60	.65	.82	.72	.....	.79	.54	.47	.67
8. Kepner B	.71	.69	.53	.56	.71	.66	.79	.....	.38	.40	.60
9. V. W. H. R. A	.49	.49	.35	.53	.56	.45	.54	.38	.....	.57	.48
10. V. W. H. R. B	.51	.45	.44	.52	.55	.44	.47	.40	.57	.....	.48

TABLE XLIV

INTERCORRELATIONS CORRECTED FOR ATTENUATION

TEST FORM	1	2	3	4	5	6	AVERAGE
1. Gregory	.....	.80	.86	.88	.91	.71	.83
2. Barr	.80	.....	.80	.74	.79	.72	.77
3. Pressey-Richards	.86	.80	.....	.82	.83	.77	.82
4. V. W. S-3	.88	.74	.82	.....	.79	.67	.78
5. Kepner	.91	.79	.83	.79	.....	.78	.82
6. V.W.H.R.	.78	.72	.77	.67	.78	.....	.74

Table XLIV shows that the lack of close relationship between the various tests is genuine rather than due to unreliability alone. It may be concluded that present history tests do not measure the same ability, but different abilities in the main, although there is, of course, a substantial degree of agreement in the functions covered. We have, therefore, additional evidence from Table XLIV supporting one of our earlier assertions to the effect that the teaching of history is as yet far from settled as to content and aim.

It was also suggested in connection with Table XLIII that the Kepner Background Test in Social Science correlated about as highly with any of the avowedly purely factual tests as the latter do with each other. Table XLIV gives additional evidence to the same effect. It is doubtful, therefore, if this test is to be regarded as essentially unlike the other tests purporting to be measures of achievement in United States history. The highest corrected coefficient reported in Table XLIV is that between the Kepner and the Gregory tests.

The Van Wagenen History Reading Tests would seem to have a clearer claim to the measurement of abilities other than pure knowledge of factual history.

Scoring and mechanical features of the tests. With the exception of the Barr tests all of the six tests studied were easy and objective of scoring. The Barr tests, however, required about four times as long as any of the others in scor-

ing, because of the unfortunate use of such devices as the drawing of lines in matching tests. This results in tangles of criss-crossed lines which are very difficult to score.

The Van Wagenen tests are easily marked, but require a somewhat complicated procedure in arriving at final scores.

The Gregory tests suffer somewhat from crowded typography in places, thus leaving insufficient space for answers. The type lines are also too long for the size and leading of the type employed. This fact, together with a poor choice of type face, makes for difficulty in reading the test items.

The Barr tests are printed on a quality of paper too transparent to be entirely satisfactory.

The Pressey-Richards test suffers most in mechanical make-up, as the quality of paper and printing is not satisfactory.

Validity of the test content. There seems to be no satisfactory way of attacking the validity of content of the six tests other than the data already given on reliability and intercorrelations. No outside criterion is available. Moreover, no adequate accounts of the validation of these tests seems to be available.

The Gregory, Pressey-Richards, and Kepner tests appear to possess about equal validity. The Barr and Van Wagenen tests appear slightly inferior to the others. The differences are slight in all cases, and none of the tests seems to be more than moderately satisfactory.

III. SUMMARY AND CONCLUSIONS

(1) History tests to date have not earned widespread acceptance.

(2) In order of decreasing reliability, the tests studied ranked as follows: Pressey-Richards, Gregory, Kepner, Van Wagenen S-3, Barr, and Van Wagenen History Reading (Table XLII).

(3) The equivalence of forms in the Gregory, Barr, Kepner, and Van Wagenen History Reading tests is close enough for all practical purposes (Table XLII).

(4) In order of increasing time required for answering the items, the tests ranked as follows: Van Wagenen S-3, Pressey-Richards, Kepner, Van Wagenen History Reading, Gregory, and Barr (Table XLII).

(5) The intercorrelations among the tests are never high. The average intercorrelation of each test against all others showed no significant differences other than in the case of the Van Wagenen History Reading Scales which gave relatively much lower intercorrelations (Table XLIII).

(6) The corrected coefficients are also never very high, showing a marked lack of genuine agreement of abilities covered by these six tests (Table XLIV).

(7) The tests are satisfactory in the mechanics of scoring, with the exception of the Barr tests.

(8) Minor mechanical defects are to be noted in certain of the tests, particularly in the quality of paper and printing in the Pressey-Richards. The reading difficulty of the Gregory tests is increased by crowded printing and a poor selection of type face. The Van Wagenen tests have a rather involved system of finding final scores.

APPENDIX

Correlation coefficients and prediction. It has been necessary in the treatment of results of the several investigations reported in the preceding chapters to make reference to coefficients of correlations (particularly reliability coefficients) in terms of the accuracy of prediction possible from coefficients of various magnitudes. Thus, in Chapter I it was asserted that a correlation between two sets of marks on the same examination papers of .62 was roughly 20 per cent better than chance assignment of marks. It is the purpose of these closing paragraphs to present in somewhat greater detail the basis for such assertions.

It is a well known fact that the accuracy of predicting the values of one variable from another variable, when the correlation between the two variables is known, is given by the formula, $\sigma_{1.2} = \sigma_1 \sqrt{1-r_{12}^2}$. This is the familiar formula for the standard error of estimate. Kelley[1] has designated the radical term of this formula as the *coefficient of alienation.*

The meaning and utility of the coefficient of alienation as applied to the treatment of examinations by correlation methods can be made clear by an illustration. First, it should be pointed out that perfect relationship or agreement is represented by a correlation of 1.00, and that absolute lack of relationship or agreement is represented by a correlation of 0.00. That the amount of relationship is not a linear function is shown by the radical term, $\sqrt{1-r^2}$, in the formula.

Table *A* shows the coefficients of alienation and the re-

[1]T. L. Kelley, *Statistical Method* (Macmillan, 1923), pp. 172-175, and elsewhere.

duction of the standard error of estimate over zero correlation in terms of percentages.[1]

TABLE *A*

COEFFICIENTS OF ALIENATION AND THE REDUCTION IN THE STANDARD ERRORS OF ESTIMATES FOR VARIOUS VALUES OF r

(*a*)	(*b*)	(*c*)
r	$\sqrt{1-r^2}$	PER CENT OF REDUCTION IN STANDARD ERROR OF ESTIMATE
.00	1.000	0.0
.10	.995	0.5
.20	.980	2.0
.30	.954	4.6
.40	.917	8.3
.50	.866	13.4
.60	.800	20.0
.70	.714	28.6
.80	.600	40.0
.866	.500	50.0
.90	.436	56.4
.95	.312	68.8
.96	.280	72.0
.97	.243	75.7
.98	.199	80.1
.99	.141	85.9
1.00	.000	100.0

Let us assume that the true marks of a class of 100 pupils are known. Let us assume further that these 100 marks are written, one by one, on gun wads such as are used in shotgun shells. If these gun wads are placed in a box and mixed thoroughly and then are drawn out one at a time by chance and assigned to the 100 pupils in the order in which they are seated in the schoolroom, we would expect the correlation of the marks thus assigned by chance to be approximately zero with the true marks on the same 100 pupils. At any rate, the departures of a series of such correlations from zero would be relatively small and would form a dis-

[1]G. M. Ruch, *The Improvement of the Written Examination* (Scott, Foresman and Co., 1924), p. 143. A similar table occurs in Kelley, *op. cit.*, p. 174.

tribution summarized by the formula for the probable error of a coefficient of correlation, viz.,

$$\text{P. E.}_r = \frac{1-r^2}{N} \cdot 6745$$

When the expression "chance assignment of marks" has been used in these pages, a situation roughly parallel to the gun-wad method was implied. When, therefore, two persons independently assigning marks to the same set of papers agree to the degree represented by a correlation of .62, it will be seen from Table *A* that this agreement is roughly 20 per cent better than chance assignment. More exactly, the agreement is 21.5 per cent better than chance. Table *A* shows further that half-of-perfect agreement requires an r of about .866. Correlations are truly high only when they are well above .95, since, at .95 the prediction possible from such an r is but 68.8 per cent better than chance.

In the treatment of the actual data of this monograph the true marks are never known. The only estimates of the probable truth with respect to the reliability of an examination come through the determination of "reliability coefficients," i. e., the correlation obtaining between two independent markings on the same set of papers, or through a second type of reliability coefficient, viz., the markings assigned by the same person to two different sets of examinations written by the same pupils. The two sets in the present investigations were usually the examinations for the two consecutive years, 1923 and 1924. It has been the convention of present discussions to brand these two types of reliability coefficients as the "reliability of scoring" and the "reliability of sampling," respectively, although these terms have been used merely in an approximate sense, especially in the case of the second type, where errors of scoring as well as of sampling enter.

Finally, it should be stated that when two competent persons read the same set of papers and arrive at different

marks for the same individuals, we have no definite knowledge of the real merits of the individual papers. If the agreement should turn out to be perfect ($r = 1.00$), the assumption would seem to be justified that both sets of marks are perfectly accurate. When the disagreement is absolute ($r = 0.00$), the only conclusion which can be drawn is that one or both sets of marks are worthless. When the agreement is greater than zero but less than unity, we are left in doubt about the real merits of the pupils' papers. Chapter II, Table I, presented cases where the reliability coefficients (of the several types) were practically zero, and also cases where the agreement approached unity. Negative correlations, from the standpoint of prediction, have the same significance as positive values. Negative r's as reliability coefficients have a somewhat different significance, since they mean that one reader *tended* to give high marks, on the whole, to pupils to whom the second reader gave low marks. The negative r's were, however, always small.

S0-FAL-593